Conceptual and computational framework for logical modelling of biological networks deregulated in diseases

Arnau Montagud, Pauline Traynard, Loredana Martignetti, Eric Bonnet, Emmanuel Barillot, Andrei Zinovyev, Laurence Calzone

OXFORD

Briefings in Bioinformatics, 20(4), 2019, 1238–1249

doi: 10.1093/bib/bbx163
Advance Access Publication Date: 8 December 2017
Paper

Conceptual and computational framework for logical modelling of biological networks deregulated in diseases

Arnau Montagud, Pauline Traynard, Loredana Martignetti, Eric Bonnet, Emmanuel Barillot, Andrei Zinovyev and Laurence Calzone

Corresponding author. Laurence Calzone, Institut Curie, PSL Research University, Mines Paris Tech, INSERM, U900, F-75005, Paris, France. Tel.: +33156246924. E-mail: laurence.calzone@curie.fr

Abstract

Mathematical models can serve as a tool to formalize biological knowledge from diverse sources, to investigate biological questions in a formal way, to test experimental hypotheses, to predict the effect of perturbations and to identify underlying mechanisms. We present a pipeline of computational tools that performs a series of analyses to explore a logical model's properties. A logical model of initiation of the metastatic process in cancer is used as a transversal example. We start by analysing the structure of the interaction network constructed from the literature or existing databases. Next, we show how to translate this network into a mathematical object, specifically a logical model, and how robustness analyses can be applied to it. We explore the visualization of the stable states, defined as specific attractors of the model, and match them to cellular fates or biological read-outs. With the different tools we present here, we explain how to assign to each solution of the model a probability and how to identify genetic interactions using mutant phenotype probabilities. Finally, we connect the model to relevant experimental data: we present how some data analyses can direct the construction of the network, and how the solutions of a mathematical model can also be compared with experimental data, with a particular focus on high-throughput data in cancer biology. A step-by-step tutorial is provided as a Supplementary Material and all models, tools and scripts are provided on an accompanying website: https://github.com/sysbio-curie/Logical_modelling_pipeline.

Key words: pipeline of tools; logical modelling; genetic interaction; robustness analysis; data integration; step-by-step tutorial

Arnau Montagud is a postdoctoral researcher in Computational Systems Biology of Cancer group at U900, Institut Curie, PSL Research University, Mines Paris Tech, INSERM. His research interests are related to Boolean modelling and data analyses in cancer biology.
Pauline Traynard was a postdoctoral researcher in Computational Systems Biology of Cancer group at U900, Institut Curie, PSL Research University, Mines Paris Tech, INSERM at the time of writing. She is now Applications Manager at Lixoft, her interests are related to model-based drug development.
Loredana Martignetti is a researcher in Computational Systems Biology of Cancer group at U900, Institut Curie, PSL Research University, Mines Paris Tech, INSERM. Her research interests are related to data integration in cancer projects.
Eric Bonnet is a researcher in Centre National de Recherche en Génomique Humaine, Institut de Biologie François Jacob, CEA. His research interests are related to computational genomics.
Emmanuel Barillot is a researcher head of the Computational Systems Biology of Cancer group and head of U900 at Institut Curie, PSL Research University, Mines Paris Tech, INSERM. His research group focuses on data analyses, Boolean modelling and knowledge formalisation in the form of maps.
Andrei Zinovyev is a researcher and scientific coordinator of Computational Systems Biology of Cancer group at U900 unit, Institut Curie, PSL Research University, Mines Paris Tech, INSERM. His research interests are data dimension reduction and managing mathematical model complexity applied to cancer biology.
Laurence Calzone is a researcher in Computational Systems Biology of Cancer group at U900, Institut Curie, PSL Research University, Mines Paris Tech, INSERM. Her research interests are Boolean modelling and data integration in cancer projects.
Submitted: 14 August 2017; **Received (in revised form):** 24 October 2017

Introduction

Signalling pathways can be defined as sets of proteins and genes that transmit environmental signals from the membrane to the nucleus. They are of utmost importance, as this is the way that the cell senses its surroundings and communicates with other cells. Moreover, these pathways are entangled into complex networks [1, 2] and can be highly deregulated in diseases such as cancer, often described as a network-alteration disease [3].

To fully understand the cellular complexity, the regulations and crosstalks between the important signalling pathways can be recapitulated in the form of a network. This network can be of several types [4], but for our study, we will focus on influence or regulatory networks (referred to as activity flow diagrams in Systems biology graphical notation (SBGN) standard format [5, 6]), where nodes represent genes, mRNAs, proteins, complexes or processes, and where edges correspond to the influence of a source onto target entities. The topology of these networks can be studied and constitutes a valuable source of information *per se*, especially when accurate annotations allow for relevant descriptions of this architecture. These networks can also be translated into mathematical models and provide more insights on the contextual regulation of the processes described and their dynamics.

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, to identify new hypotheses and to test experimental hypotheses and predict their outcome. In short, a mathematical model can help reason on a problem.

Discrete models (stemming from the logical models) are becoming more popular for exploring cell fate decisions, or particular dysfunctions in biological processes [7–12]. Logical models are particularly appropriate when the question is qualitative, e.g. which genomic alterations can lead to an increase of cell proliferation and which genomic alterations need to be combined to cause resistance to some particular drugs or combinations thereof. In biomedical literature, there are few time-resolved data on the detailed dynamics of molecular processes of tumours, as well as reported kinetic rates of the reactions, but above all, there is frequently only approximate understanding of the precise biochemical reactions that are involved in the malfunctioning of the normal cell. Logical modelling is an approach, which is abstract enough to work under these constraints and still can deliver interesting insights into the systems' behaviour. This formalism has already been fruitful to identify the role of molecular entities in the cellular response to various perturbations, such as mutations, in the context of cancer [7, 13–17] but also in T-cell differentiation [18], in development in *Drosophila* [19] among others.

We propose a framework of computational methods that we have developed to answer biological questions using logical formalism (Figure 1). Overall, the set of tools presented here builds up a pipeline that allows users to characterize and summarize properties of a logical model by following simple and automatic procedures. We also present different ways to knit data on top of the interaction network as a way to understand system-wide effects, or to understand the data through the model.

Methodology

In the present work, we apply a set of tools that characterize a mathematical model oriented towards the study of phenotypes and their regulation. These tools can be applied to any logical model stored in the standard SBML qual format [20]. As a working example, we use a previously published logical model focusing on pathways leading to the early steps of the metastatic process [13].

A detailed account of all the steps followed to obtain the results presented in this work can be found in Supplementary File 1. We also provide another example on a gastric cancer logical model [21] in the models' folder of the GitHub repository. All models used to illustrate the pipeline, tools and scripts, together with a Dockerfile (to use them in a Docker container) are provided on an accompanying website at the following address: https://github.com/sysbio-curie/Logical_modelling_pipeline. A stewardship script (Stew.sh) can be used to automatize the different sections of the pipeline and retrieve basic plots with the most probable phenotypes. It is meant to be used together with the docker container and requires minimum information from user (model files and definition of input and output nodes). This

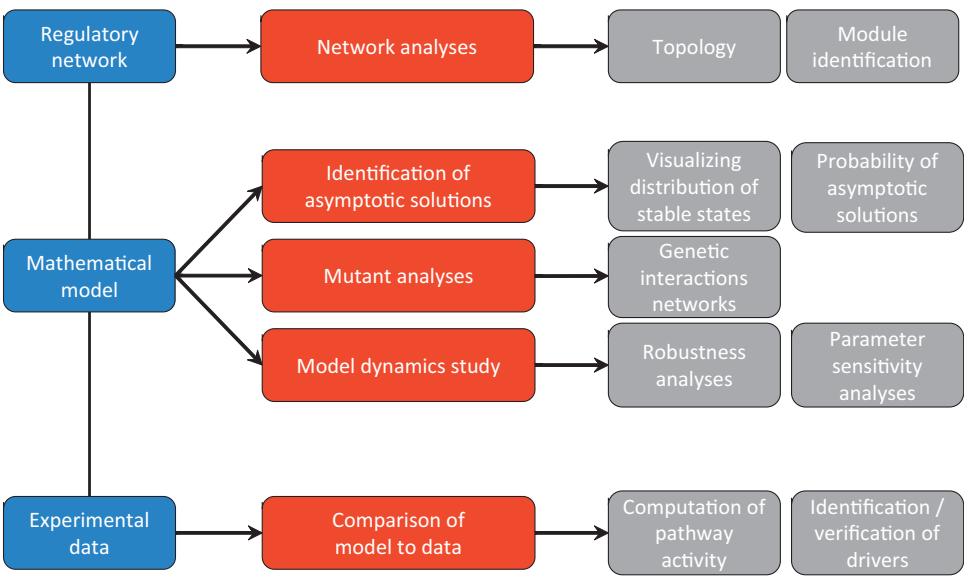

Figure 1. Modelling pipeline. Some tasks (in red, central column) can be performed at different levels (in blue, left column): on the regulatory network, the mathematical model or using experimental data. The results of the possible tasks (in grey, right column) can take several forms.

Table 1. Tools used in present logical modelling framework

Pipeline part	Tools used	Other tools that could be used
Constructing the model network	Cytoscape [22], GINsim [23], Databases	CellDesigner [24], SIGNOR [25], OmniPath [26], PHONEMeS [27]
Translation of the network into a mathematical model	GINsim, MaBoSS [28]	BoolNet [29], CellNOpt [47]
Identification and analyses of asymptotic solutions	GINsim, MaBoSS, R [31]	BoolNet, EpiLog (http://epilog-tool.org/), Gene Network Analyzer [32], PyBoolNet [33], SQUAD [34], AVATAR [35], CellNOpt, Pint [72]
Model reduction	BiNoM [36], GINsim	
Mutant analyses	BiNoM, MaBoSS, R, ViDaExpert [37]	GINsim, BoolNet, EpiLog, Gene Network Analyzer, PyBoolNet, SQUAD, AVATAR, CellNOpt
Robustness analysis of logical gates	BiNoM, MaBoSS, R, ViDaExpert	SQUAD
Using the model as a scaffold for data integration	ROMA [38]	NaviCell [39], GSEA [40], network inference methods (for review, [41])
Using data as priors of model construction	Lemon-Tree [42], ROMA	CLR [43], IRWRLDA [44], PBMDA [45]

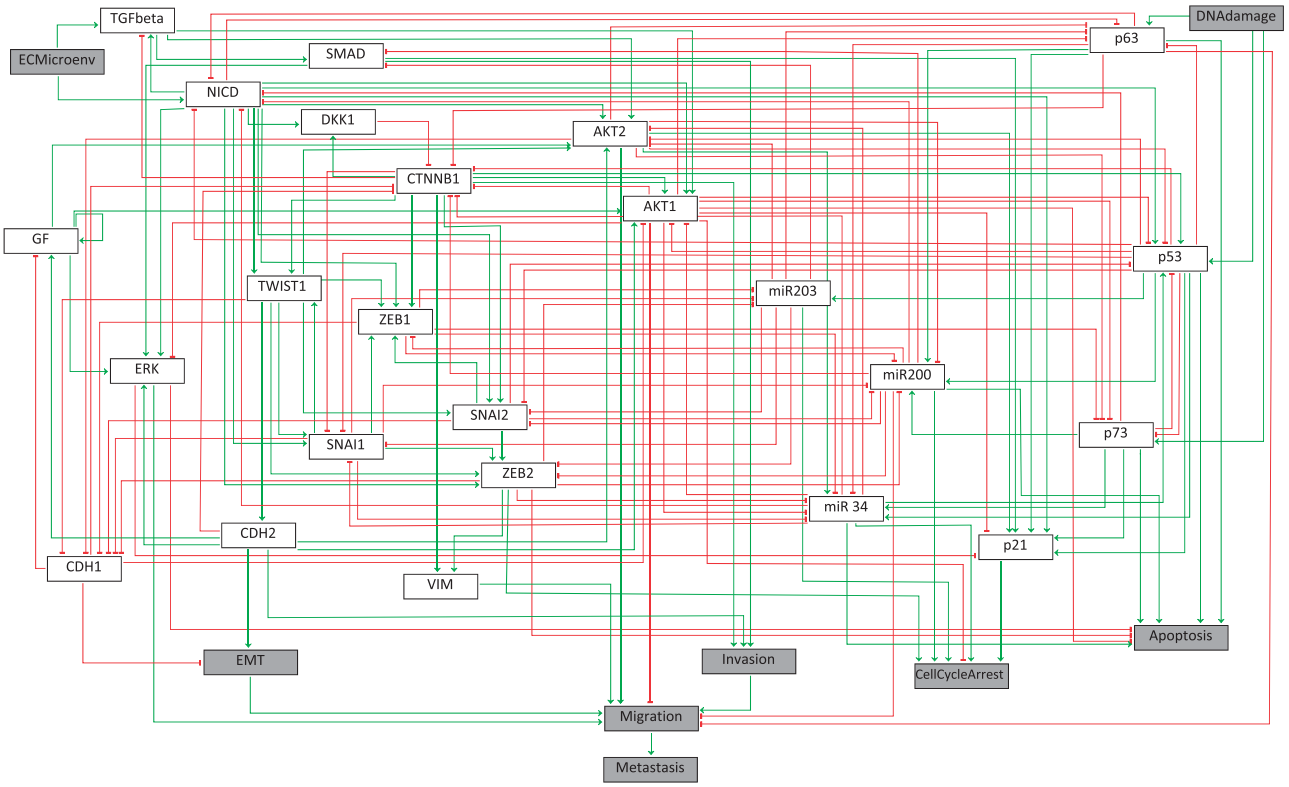

Figure 2. Influence network of the metastasis process taken from Cohen *et al.* [13].

script can be found in 'run all analyses' folder of the GitHub repository. We have detailed the tools used in the present pipeline in Table 1, as well as other tools that could perform the different tasks for the sake of completeness. Additionally, we have done performance tests of the different sections of the pipeline in Unix, Windows and Mac systems (Supplementary Table S1 in the doc folder of the GitHub repository).

The regulatory network

The construction of a regulatory network that recapitulates the processes participating in the biological study should start from a clear and precise question. The formulation of this question then leads to identifying which molecular entities need to be included to characterize these processes, how much detail is needed to answer the question and how many signalling pathways should be described. Once this is done, we can study how the nodes in the network affect each other, what the crosstalks between the studied pathways are and how phenotypes depend on the activity of these pathways.

In the example used here, the question focuses on identifying necessary pathways for the commitment of the cell to a pro-metastatic phenotype. The model describes mechanisms of early steps of tumour invasion, by showing epithelial-to-

mesenchymal transition, by invading surrounding tissues, and/or by escaping apoptosis. The model was initially built from disseminated sources of information gathered into a network. In this particular case, nodes represent genes, proteins, complexes or molecular processes, and edges are positive or negative influences. The influence network was then translated into a logical model. The model we use as an example comprises 4 input nodes (ECMicroenv, DNAdamage, GF and TGFbeta), 6 output nodes (EMT, Invasion, CellCycleArrest, Apoptosis, Migration and Metastasis), 24 internal nodes and 157 interactions (Figure 2). More details can be found at [13].

The mathematical model

Translating the network into a mathematical model

A mathematical model can apprehend the biological problems that are formulated in an exact and unambiguous language, on some counter-intuitive observations or help explain their causes. The regulatory network is converted into a mathematical model by associating some mathematical terms to each node of the network. This translation needs to reproduce the expected dynamics and comply to the possible constraints of the model (phenotypes of known mutations, or reported experimental conditions). There exist a variety of mathematical formalisms to be used and the choice depends on the scope of the question and the types of data available. Some reviews already expose panels of existing formalisms and computational simulation methods and can help in choosing the best mathematical formalism for a particular case [4].

We chose to focus on logical (also termed discrete and Boolean) modelling. A brief glossary of the terms used in logical formalism can be found in Supplementary File 2. Logical modelling is best appropriate when the questions are qualitative, when the experimental data are discrete, when there is limited information about the reactions rates or when the detailed mechanisms of the biochemical reactions are poorly known. Despite scarcity or discreteness of data, logical modelling gives an array of qualitative results describing various perturbations on the model structure or dynamics that are of interest for the scope of our question.

To explore and simulate the behaviours of logical models, one can find a handful of different software, e.g. BoolNet [29], GINsim [23], SQUAD [46], CellNOpt [47], MaBoSS [28, 48] and Avatar [35]. These tools are able to communicate, as most of them directly support the standard format SBML qual [20]. We focus on two of them: first, GINsim can be used for the friendliness of the interface and long-term support of software development. GINsim can easily inform on all stable states of the model, the functionality of positive and negative circuits or propose reduced models. MaBoSS software has been developed to perform stochastic simulations on the logical model and offers a more quantitative outcome of the asymptotic solutions, as shown below.

For the case of our example (referred to as 'the metastasis model'), the first tasks consist in validating that the network model accounts for what is known about the regulation processes leading to metastases, and thus, that the structure of the network is coherent with published experimental observations and the data that could be gathered. A logical rule was associated with each node of the network according to the available information. If rules could not be deduced from the literature, they were defined to fit all the constraints that the model should comply with. These constraints can be experimental results and are usually qualitative, of the type 'mutant of gene *A* has a reduced apoptosis' or 'expression of gene *B* is increased'. This model verification can be traced on Cohen *et al.* [13] Supplementary Material and are not detailed here.

Analysing asymptotic solutions

Given different sets of inputs, the system exhibits different sets of solutions, which, in a logical model, are called attractors. They can be fixed points, or stable state solutions, corresponding to a final state in the state transition graph, which recapitulates the dynamics of the model. They can be complex attractors, or limit cycles, when a set of model states cycles with no outgoing transition in the state transition graph; thus, there is no escape from it (see Supplementary File 2 for definition of these terms).

In the metastasis model, nine stable states are identified. These stable states can be grouped into phenotypes: four of them are related to Apoptosis commitment, two of them to EMT, two of them to Metastasis with EMT and the remaining one is Homeostasis, where all nodes have 0 value, except for an internal node: CDH1. These results can be studied in Supplementary File 1 and their biological conclusions can be found in [13].

One way to visualize the solutions of the model is to reduce the dimension of the table recapitulating the stable state values. The simplest approach consists in applying the principal component analysis (PCA) on the collection of all stable states (Figure 3). PCA plots allow the user to group the stable state solutions into clusters and determine which variables contribute the most to the cluster characterization. In Figure 3, EMT stable states (FP6 and FP7) 'jump' to Metastasis (FP8 and FP9) when TGFbeta is activated (or downstream members of the TGFbeta pathway, such as SMAD and DKK1), highlighting the role of the TGFbeta pathway in triggering metastasis. This analysis is insightful, for instance, to find the underlying model's mechanisms and to quickly spot dependencies among phenotypes. Note that in this analysis, complex attractors such as limit cycle solutions are not considered. The identification of complex attractors is harder in large models and is currently the object of methodological studies [49]. To produce the PCA plots, several packages in R software such as FactoMineR [50] can be used.

In addition to studying what the stable states of the model are and characterizing them, some questions can arise on the reachability conditions for these stable states, e.g. if transient effects can be observed and play a role in cell fate decision and if the model solutions are robust to small perturbations. For this purpose, MaBoSS [28, 48], a C++ software for stochastically simulating continuous/discrete time Markov processes defined on the state transition graph describing the dynamics of a logical model, can be used. In MaBoSS framework, the rates up (change from OFF to ON) and down (from ON to OFF) for each variable of the model can be explicitly defined to represent physical kinetic rates of the variables' turnover [48]. Probabilities to reach a phenotype are thus computed by simulating random walks on the probabilistic state transition graph. In this example, the outputs of MaBoSS focused on the read-outs of the model, but it can be done for any node of a model. The nine stable states are assigned one of the four 'meta-phenotypes': Apoptosis, EMT, Metastasis and Homeostasis. In the Apoptosis phenotype, the node Apoptosis is 1; in the EMT phenotype, EMT is 1 and Metastasis is 0; in Metastasis phenotype, the nodes EMT, Invasion, Migration and Metastasis are 1; and for the

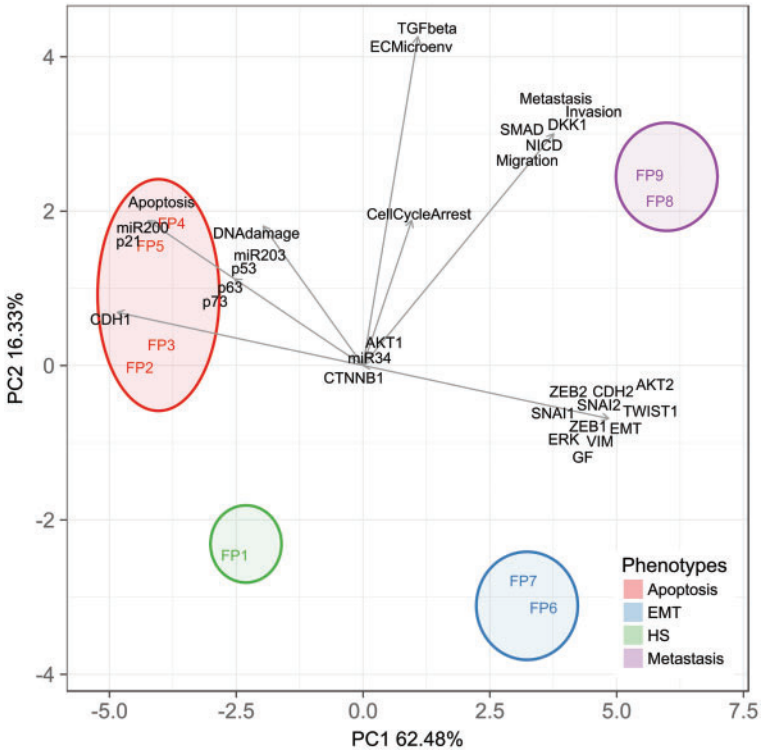

Figure 3. PCA bi-plot of stable state solutions of the metastasis model. The nine fixed points are represented here: FP1 corresponds to the homeostatic stable state (HS), FP2–FP4 to the apoptotic stable states, FP6 and FP7 to the EMT stable states and FP8 and FP9 to the metastatic stable states. The arrows show the directions of the contributions. PC1 shows that the EMT regulators contribute the most to the EMT and Metastasis stable states, whereas PC2 shows that TGFbeta pathway promotes Migration and thus Metastasis.

Homeostasis phenotype, only an internal node, CDH1 is activated in this stable state.

For each simulation in MaBoSS, some initial conditions are defined and a maximum time is set to ensure that the simulations reach asymptotic solutions (Supplementary File 2). There are two ways to visualize the results: (1) the trajectories for particular model states (states of nodes) can be interpreted as the evolution of a cell population as a function of time (Figure 4A); that way, transient effects could be highlighted. Alternatively, (2) asymptotic solutions can be represented as pie charts to illustrate the proportions of cells in particular model states. This representation is particularly handy when two cell conditions (e.g. altered environment conditions or component perturbations) are compared and the proportions of the model states change from one condition to another (Figure 4B and C).

Simplifying the model structure

When models are too complex (with a high number of variables), it becomes difficult to get insight on which molecular mechanisms described by the network are responsible for which behaviours or simply, to simulate the model. It is important to know that this reduction will come at the cost of losing details and granularity of the model. This trade-off is usually accepted by the modellers. One solution is to reduce the model to a small number of variables while maintaining the number of solutions of the initial comprehensive model.

There are several ways to reduce a logical network model: among them, masking nodes or lumping nodes into modules. For the former, GINsim [23] allows selecting components for reduction; their regulators then regulate their targets whose logical rules are appropriately modified. This reduction maintains the stable states. This may be interesting when some parts of the model can be hidden to focus elsewhere, or when different parts of the model have different temporal scales (the reduction considers that the reduced components are faster). Alternatively, Cytoscape plugin BiNoM [39, 51] helps the modeller in the reduction of the regulatory network by lumping nodes into modules, focusing solely on the structure of the network and not the logical rules. The modules become nodes of the model, and should represent the global behaviour of the nodes inside the module. So far, the inference of the signs of the edges and the assignment of the logical rules of the reduced model need to be done manually. This is usually a complex task as the solutions of the reduced model and the original model need to be the same. The approaches of GINsim and BiNoM answer different types of questions: the first one maintains the dynamics of the complete model but concentrates on some players (as can be seen in [14]), whereas the second one highlights mechanisms or motifs (modules) that generate particular response behaviours (as can be seen in [3]). Both methodologies are showcased in Supplementary File 1, where detailed step-by-step procedures can be found.

Analysing mutants

The construction of a logical model involves a step of model validation that is the verification that the model is coherent with known facts or experiments. In this section, we show how a priori knowledge is used as constraints the model must comply. These constraints can be qualitative information related to mutants of genes of the model (in the form of statements like '*TP53* deleted mutant in mice leads to reduction of apoptosis'),

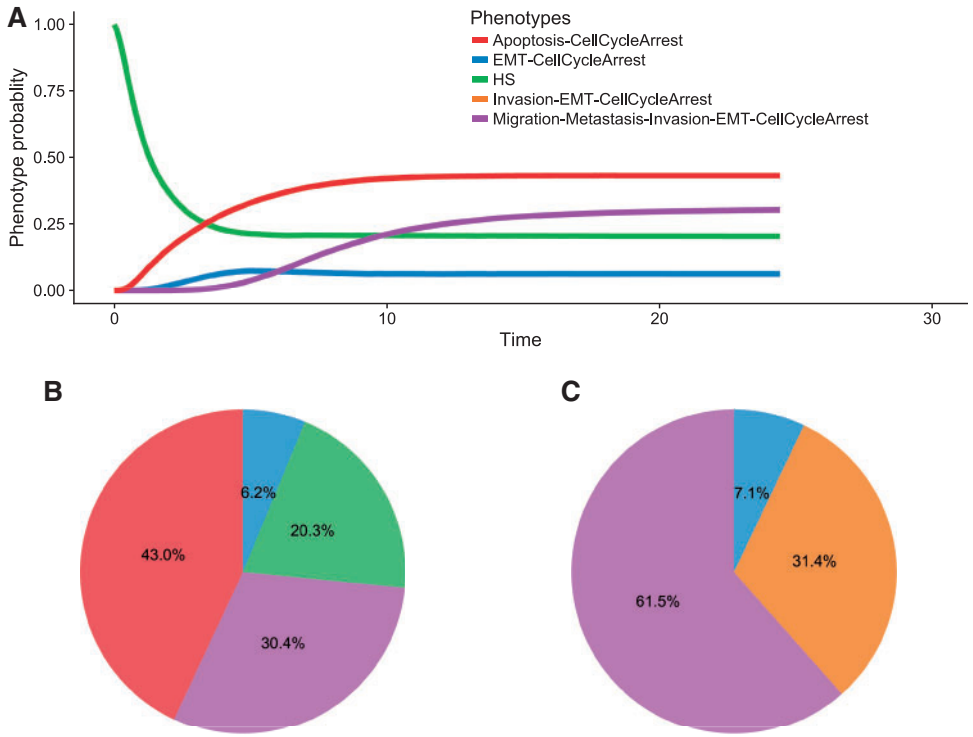

Figure 4. Plot of the metastasis asymptotic solutions using MaBoSS. Initial conditions were chosen such that: all internal nodes had initial values set to 0 and inputs nodes (ECMicroenv, DNAdamage, GF and TGFbeta) were randomly set to 0 or 1. (**A**). Evolution of the probabilities of read-outs of the model (phenotypes), based on 50 000 trajectories; (**B**) Pie chart of asymptotic solutions for the wild-type case; (**C**) Pie chart of asymptotic solutions for the NICD overexpressed mutant. EMT stands for Epithelial to mesenchymal transition, HS stands for Homeostatic state.

or conditions for experiments ('cells in response to TGFbeta have increased EMT'). Simulating these constraints consists in modifying the logical rule or node's value of the genes related to the experiment. A *TP53*-deleted mutant will be simulated by setting the activity of the node p53 to 0, ignoring the initial rule of p53. In the case of a *TP53* mutant where experiments would report a reduction in apoptosis [52], the model states related to apoptosis (the output node Apoptosis in the example) should exhibit a diminished probability when compared with the wild type. The model needs to comply with all (or most of) the constraints to claim that the model is coherent with the known biological facts. For the mutants that have not been experimentally performed yet, the model solutions can be treated as predictions of their behaviour.

We provide a set of scripts that computes and simulates all single (one gene altered in one cell) and double mutants (two genes altered in one single cell) of any logical model using MaBoSS. Our methodology quickly determines probabilities for all mutants' phenotypes. Mutant probabilities can be compared with the wild type to see the extent of the different phenotype probabilities' shifts. Using MaBoSS, one can study and compare quantitatively the effects of mutants on a given phenotype such as Metastasis. Combinations of mutations are of particular interest, as they can be used to conclude if two alterations are, for instance, synergistic (the alteration of the double mutant has more effect on a phenotype than the sum of the single alterations) or synthetic lethal (the double mutant is not viable, while single mutants are viable). These terms are different cases of epistasis behaviours: when the effect of one gene in the phenotype is either not modified or increased in the presence of another genetic alteration.

Predicting genetic interactions

The mutant probabilities obtained with MaBoSS in the previous step can be used to analyse the effect of mutations on double mutants. This epistasis study explores the combined effects of all double mutations in comparison with wild type and single mutants [53]. The method was applied to the metastasis model [13, 53] with respect to all outputs: Homeostasis, EMT, Migration, Metastasis, Apoptosis and Cell Cycle Arrest (Figure 5).

The Metastasis phenotype revealed to be the one showing the highest deviation from the wild-type probability: 45.77% of the mutant combinations abolish this phenotype probability to 0 (Supplementary Figure S4 in Supplementary File 1). In fact, these 938 mutants are mainly combinations of knockouts of genes that are necessary for the activation of the Migration node (which is the node whose activation depends on the highest number of internal nodes) as well as overexpression of genes that are anti-migratory or pro-apoptotic. These genes can be considered as candidates for drug-targeted therapies, as they might considerably reduce the occurrence of metastases.

Additionally, double alterations in some pairs of genes are predicted to increase significantly the triggering of metastasis, which highlights alterations, or mechanisms, that should be carefully monitored in cancer patients, as they may have a role in increasing metastatic predisposition that could lead to aggravated cancer condition. These alterations were gains of functions of three genes: *AKT2*, *SNAI1* and *TWIST1* that, indeed, correspond to patients with bad prognosis [54]. In Figure 5, *AKT2*, *SNAI1* and *TWIST1* gains of function are the single mutants that move the wild-type state the most towards Metastasis and EMT phenotypes (AKT2_oe, SNAI1_oe and TWIST1_oe, dark red dots). We

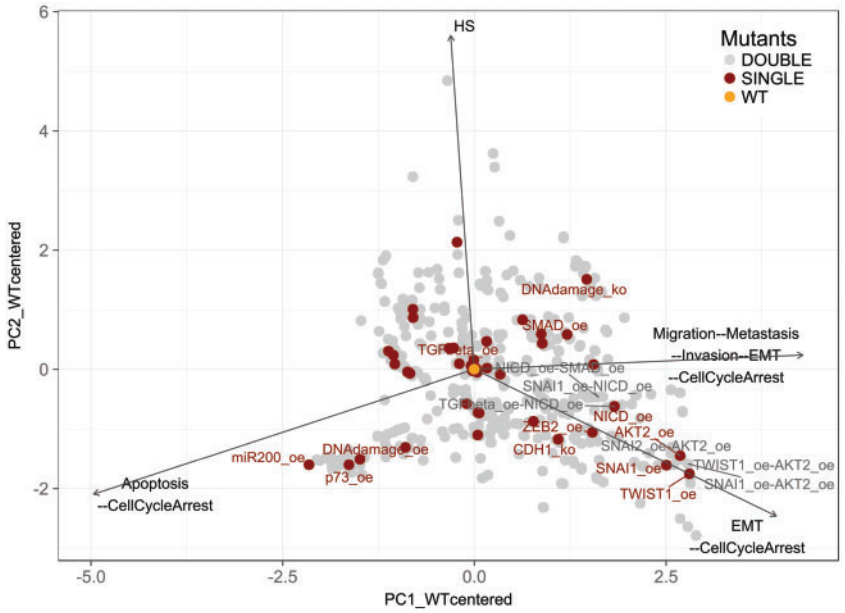

Figure 5. PCA bi-plot of the probability profiles over the set of four studied meta-phenotypes. The probabilities of all overexpressed (oe) or knockout (ko) single (red) and double (grey) mutants are plotted with the wild type (WT—yellow) as the centre of origin. The arrows show the contribution of the four phenotypes to each single or double mutant.

can also identify that the three epistatic pairs that have more effect on these two phenotypes are the combinations of *AKT2* gain of function with *SNAI1, SNAI2* and *TWIST1* gains of functions (grey dots with labels). We further discuss the use of these PCA figures in Supplementary File 1.

Robustness analysis of the logical model

The logical rules are often written such that the global behaviour is in accordance with reported facts in both the wild type and the perturbed conditions (mutations, drug treatments, etc.). In a logical framework, it is often difficult to justify the choice of the logical rules biologically and to ensure that the obtained model generates compatible dynamics with what is known. Also, we expect that there exists more than one set of logical rules that can comply with the model constraints, but some logical rules need to be written more carefully than others because they play an important role in reaching a given phenotype. One way to address this issue is to modify the logical rules and check how the solutions vary from those of the initial model. There are more than one way to do it. Hereby, we propose to modify automatically one operator at a time (from AND to OR and vice versa) and compute how different the perturbed model solutions are compared with the wild-type solutions, either at the level of the phenotype probabilities, or at the level of stable state solutions. The model can be robust with respect to either aspect when a threshold for assessing robustness is defined and met.

For the example presented here, a phenotype is considered to be robust when >50% of the (logical rules) model variants have the same probability than the wild-type model. We define a model variant as a model for which one logical operator was changed when compared with the initial model. Using MaBoSS, only two types of rule modifications are explored: one or two operators (AND in OR and vice versa) are changed per rule. Note that the provided scripts can also perform three changes: one operator in one rule, two operators in one rule or one operator in two different rules. More changes can be made but the computations might become heavy then because of a combinatorial number of possibilities. It was concluded that, for the metastasis model, all phenotypes were robust to changes except for Homeostasis. This is because of the fact that Homeostasis is a phenotype that is active when no input is present. Thus, all changes that end up activating a pathway will cause an alteration in Homeostasis activation.

In fact, looking at the combined meta-phenotype identified as the Metastasis phenotype (equivalent to the nodes Migration/Metastasis/Invasion/EMT/CellCycleArrest ON), 66.4% of the model (logical) variants have the same probability to reach this phenotype than the wild type (Figure 6), whereas for the Homeostasis phenotype, 22% of model variants have the same probability as the wild type (Supplementary Figure S1).

This robustness analysis can highlight 'weak' rules in the regulation of genes whose alteration suppresses the Metastasis phenotype. These alterations are found in the logical rules of nodes such as AKT1 (by far the biggest contributor) and p53, and they represent 9.2% of the total model variants.

Another possibility of interpretation is to check the robustness with respect to the stable states. For that, a distance from the stable states of the variant of the model to the closest stable states of the wild type is calculated. This distance corresponds to the minimal number of changes in the node activities (with 0 or 1 values) between all the stable states of the reference wild-type model and each stable state of a variant model. This computed distance, called Hamming distance, is a measurement of the perturbation suffered by the model variant on the logical operator change. This way, it is possible to identify which stable states are most robust to logical operator's perturbations and, thus, which rules should be considered a priority to check when simulations do not tally experimental data. Further discussion on these results can be found in Supplementary File 1 and in Supplementary Figure S2.

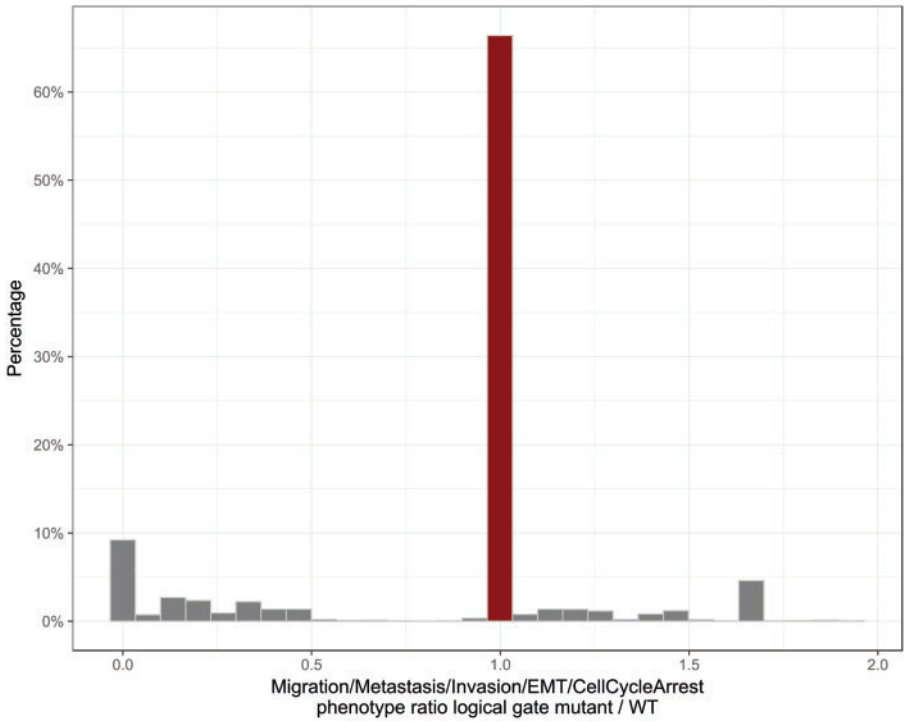

Figure 6. Robustness analysis. A distribution of ratio between Migration/Metastasis/Invasion/EMT/CellCycleArrest phenotype probability to wild-type probability. The bin centred at the wild-type value has been marked with dark red colour.

Data to model and model to data

We can further use our model to compare its results with experimental data. This allows the use of mathematical models as a tool to understand data in a more insightful way [13, 47, 52, 55, 56]. Conclusions from experiments can be interpreted under the light of the model, and unveil mechanisms that contribute to a disease state. As an illustration, we chose a data set of eight colon cancer patients treated with cetuximab: four that responded to the treatment and four that did not respond. We explored these data using the metastasis model used throughout the present work. This data set can be found at Gene Expression Omnibus under the code GEO: GSE56386 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi? acc=GSE56386).

Mapping omics data on gene regulatory networks

The visualization of data on top of a network can be informative. It is possible to show the expression of a particular protein, or gene, on the network and get some insight about the activity of its neighbours, its mechanism or the whole pathway. Taking transcriptomics data as an example, we mapped data onto the network to easily visualize the differential gene expression between two conditions, such as responders versus non-responders. Nodes that represent a gene are associated to a gene from the data set (HUGO name). The mean expression value of the two groups for each gene of the network is computed and the expression is then mapped onto the influence network using Cytoscape (Supplementary Figure S9 in Supplementary File 1). Focusing, for instance, on the EMT regulator module, it is difficult to conclude on the activation of EMT or not, as some transcription factors, the main players of the EMT response, show differences in expression and others do not. We conclude that the expression of individual genes does not provide any insight about the activity of the process.

Another approach consists in considering the activity of a gene set rather than individual genes. For the case of a transcription factor, the expression of the target genes can inform on its activity more accurately than the expression of the transcription factor itself. For the case of signalling pathways, the expression of the genes that compose the pathway can also account for its activity. Using ROMA (Representation Of Module Activity) [38], a score based on a weighted sum of the expression of the genes that compose the module can be associated. For this analysis, the modular (or reduced) network is appropriate. As mentioned above, for the transcription factors of the EMT module, for example, the gene sets correspond to their target genes, and for some pathways such as Notch or Wnt, the genes that participate in the pathway are considered, according to pathway databases, such as KEGG [57], Reactome [58] or ACSN [59] (more details on methodology on Supplementary File 1).

In Figure 7, the mean activity of each module and for each group of patients, responders and non-responders, is mapped onto the network. EMT node appears more active in non-responders than responders, which corresponds to what is expected from resistance mechanisms. In the responder group, p53 and microRNA (miRNA) nodes seem to have a higher activity, thus able to trigger Apoptosis, whereas in the non-responder group, stronger activation of the Extracellular microenvironment, TGFbeta, Notch pathways and Akt1 is observed. Interestingly, the E-cadherin module is more activated in non-responders, which is a mechanism that needs to be further explored. The significance of these differences of activity between two groups can also be statistically assessed (by computing *P*-values in ROMA). The gene sets used for this analysis

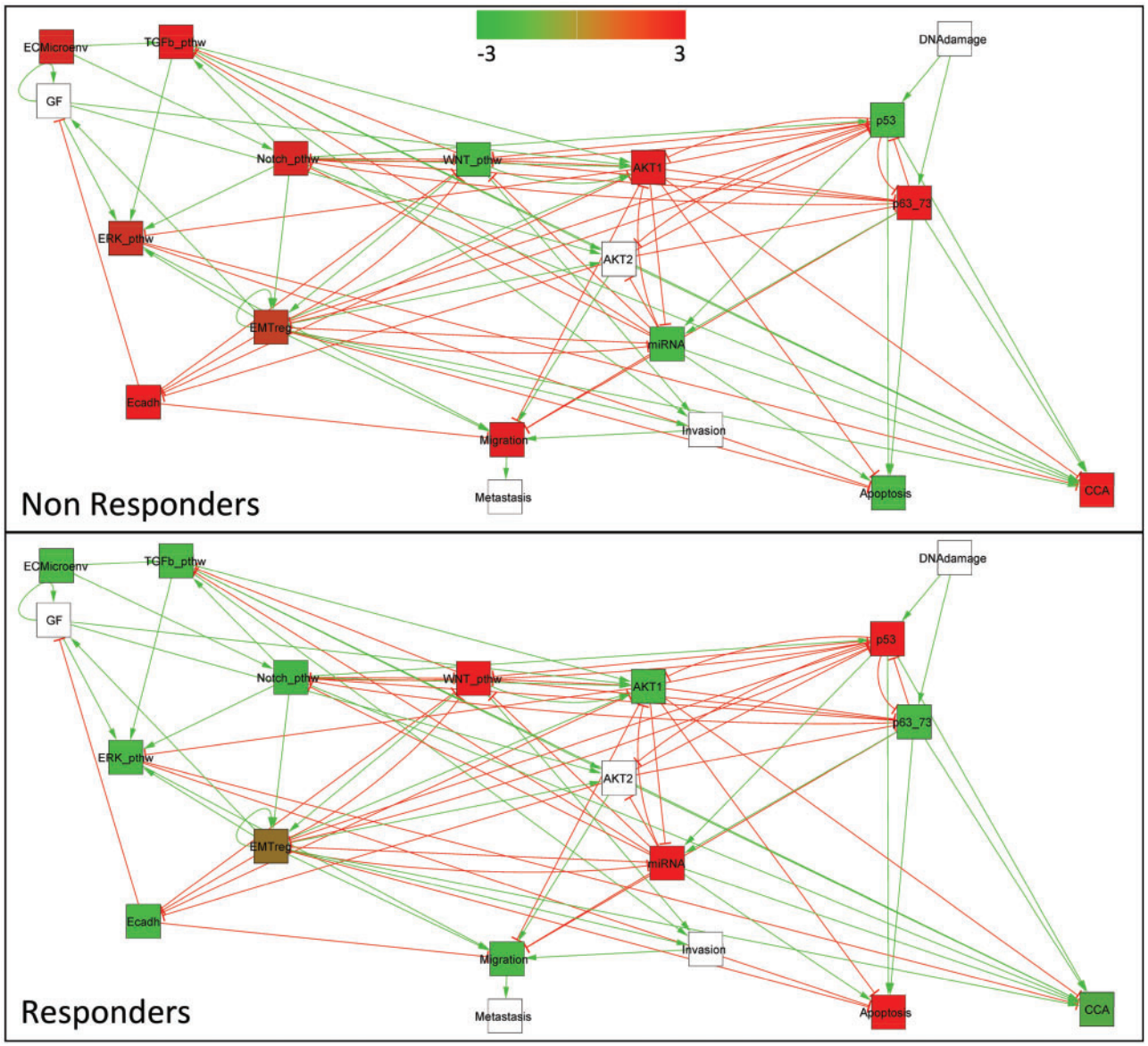

Figure 7. Gene expression data mapped onto an influence network composed of functional modules for the two groups of patients, the non-responders (upper panel) and the responders (bottom panel).

were described in the initial publication and are provided in the GitHub.

Using data for selecting appropriate components

Data exploration can also be used as a prior step to guide the model construction, and to identify relevant pathways in high-throughput data or highly variable sets of genes. Before constructing the network that would eventually be converted into a model, the data can be explored to search for appropriate components that need to be included in the model and that will permit some non-intuitive predictions.

We could start with a statistical test identifying the most differentially expressed genes, but the list might end up being too long and linking them might lead to tedious manual curation. Rather than working on individual genes, we propose to explore signalling pathways that are specific to the cancer model we wish to study or to the data set itself.

We present here two tools that we have used as means to identify pathways and mechanisms that should be included if we wanted to have such a data-specific model. We choose the same data set used for ROMA analysis (GEO: GSE56386). We can explore the content of pathway databases to find significant pathways that can represent these data. In MSigDB [40] (http://software.broadinstitute.org/gsea/msigdb), for instance, lists of pathways can be found. According to ROMA results, the modules that appear significantly over- or under-activated can be considered as candidates for an initial list of pathways (and thus candidate genes or proteins) that would need to be included in the model to fully capture the data. ROMA was applied to the data set of eight colon cancer patients to compute the activity score of gene sets downloaded from the KEGG database [57]. Among the top significant differentially activated pathways between the two groups of patients, some new ones were identified as potentially involved in the disease progression, such as pathways related to the immune response [60, 61].

Including these pathways in the network would tailor the model to these specific data.

To build a network tailored to the data, network inference methods have been widely applied. The idea is to exploit high-throughput data sources to infer regulatory relationships between genes and thus reconstruct a network from experimental data. A wide range of network inference methods have been developed [41, 62, 63] and successfully applied to various biological problems [43, 64–69]. Similarly, tools, such as the context likelihood of relatedness (CLR, [43]) algorithm, can also be used to identify interactors.

Network inference can thus be used in the context of building a logical model to prioritize the genes to be included. Lemon-Tree is one of these software frameworks dedicated to module network inference [42]. The tool can be used to assign and prioritize candidate regulators to module of co-expressed genes and thus determine which entity has an essential role and should be kept or could be otherwise discarded. The list of candidate regulators is built by using prior independent information, such as ontologies describing the biological function of a gene. In the example here, the initial list of genes from the first version of the network was used as input for Lemon-Tree together with the expression profiles of eight colon cancer patients. The results suggest that top regulator genes, such as *CDC42SE1, MKNK1* and *FGFR3*, should be part of the model to build a specific colon cancer model that explains differential cetuximab response.

These aforementioned analyses aim at listing important genes. Tools such as SIGNOR [25] or more generally OmniPath [26] can then be used to complete the network. Omipath searches in existing databases links between genes to have a comprehensive and cohesive network. Further analyses can be performed using the network structure: one can extract the minimal cut sets [70, 71] or predict association of miRNAs [45] or long non-coding RNAs (lncRNAs) [44] to diseases.

Conclusions

In the present work, we have showcased how to extract information from an existing logical model. Through a series of analyses on a model of the early steps of metastases, we have exposed different methods and tools that could be used as predictions. Many of the functionalities presented here use the outputs of the modelling tool MaBoSS, which stands as a way to fill the gap between qualitative and quantitative modelling. It is based on continuous time Markov process applied on a Boolean state space in which we explicitly specify the transition rates for each node to describe the temporal evolution of the biological process we wish to model. That way, transient effects are represented by the dynamics probability distributions, defined with a physical continuous time (unlike the standard Boolean approach where the time is discrete). The results of MaBoSS are interpretable in terms of cell population dynamics.

We have recommended several methods that can exploit the data to parameterize the model, to ensure that the most important pathways and genes are included in the model, as well as to verify the coherence of the model. In the example used in the present article, the model that describes early tumour invasion steps was built out of current knowledge to be generic, and the data were chosen to illustrate the methods. The model successfully explains colon cancer data if these are bundled in functional modules. Furthermore, if we wish to apply this model to a particular cancer, the model would need to be adapted, specified and extended. In this case, the model expansion depends on the data, and we suggest some methods to identify relevant pathways and genes, such as ROMA and Lemon-Tree that.

Future perspectives of this work are to expand this pipeline with other existing tools to include other analyses and studies and by doing so, to be able to improve the current results of models. For instance, tools such as BoolNet [29] or CellNOpt [47] could be easily integrated in the analysis of asymptotic solutions section of the pipeline, and tools such as AVATAR [35] or PyBoolNet [33] could be also used to study mutants of the models. Additionally, in a longer term, we plan to devise a graphical user interface that performs all the steps detailed in Supplementary File 1 to further ease the use of this pipeline. Our vision is that the present pipeline would be used as a benchmarking routine for published models, allowing for the comparison of models that have similar underlying scientific questions and the spread of its uses. To this end, this work is a contribution to the community-effort of the CoLoMoTo consortium (http://colomoto.org/) of enabling exchange and reusability of logical models for a variety of tools developed in the logical modellers' community.

We propose a pipeline of methods and resources that takes the user from a logical model to data integration and model simulations. This pipeline can be of used by novices as well as experienced modellers that are looking for a streamlined way of characterizing their biological system of interest. Our pipeline's step-by-step procedures can be followed from Supplementary File 1. The pipeline was applied to another example of multivalued model of gastric cancer and is provided as a Supplementary File. All data, scripts and examples can be downloaded from https://github.com/sysbio-curie/Logical_modelling_pipeline.

Key Points

- A wide range of free tools and resources for the study of genes' regulation networks have become available in recent years. These tools can be used in a streamlined manner to perform different kind of analyses.
- Tools presented here cover a wide range of studies that can be performed on regulatory networks, from model building to omics data analysis.
- We provide scripts and software access, so that the community can benefit from this pipeline.

Supplementary Data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Acknowledgements

The authors would like to thank Claudine Chaouiya for critical reading of the manuscript and Aurélien Naldi and Gautier Stoll for fruitful discussions.

Funding

This work has received support under the program «Investissements d'Avenir» launched by the French Government and implemented by ANR with project ABS4NGS (grant number ANR-11-BINF-0001). This work has also been partially funded by INVADE grant from ITMO

Cancer (Call Systems Biology 2012). This work has also received funding from ANR-FNR project 'AlgoReCell' (grant number ANR-16-CE12-0034). Finally, this work has also been partially funded by the ERACoSysMed research programme, which is a transnational R&D programme jointly funded by national funding organizations within the framework of the ERA-NET ERACoSysMed.

References

1. Prahallad A, Bernards R. Opportunities and challenges provided by crosstalk between signalling pathways in cancer. *Oncogene* 2016;**35**(9):1073–9.
2. Vert G, Chory J. Crosstalk in cellular signaling: background noise or the real thing? *Dev Cell* 2011;**21**(6):985–91.
3. Barillot E, Calzone L, Hupe P, *et al*. *Computational Systems Biology of Cancer*. Boca Raton, FL: CRC Press, 2012.
4. Le Novère N. Quantitative and logic modelling of molecular and gene networks. *Nat Rev Genet* 2015;**16**:146–58.
5. Mi H, Schreiber F, Moodie S, *et al*. Systems biology graphical notation: activity flow language level 1 version 1.2. *J Integr Bioinforma* 2015;**12**:265.
6. Le Novère N, Hucka M, Mi H, *et al*. The systems biology graphical notation. *Nat Biotechnol* 2009;**27**:735–41.
7. Calzone L, Tournier L, Fourquet S, *et al*. Mathematical modelling of cell-fate decision in response to death receptor engagement. *PLoS Comput Biol* 2010;**6**(3):e1000702.
8. Rodríguez A, Sosa D, Torres L, *et al*. A Boolean network model of the FA/BRCA pathway. *Bioinformatics* 2012;**28**(6):858–66.
9. Kazemzadeh L, Cvijovic M, Petranovic D. Boolean model of yeast apoptosis as a tool to study yeast and human apoptotic regulations. *Front Physiol* 2012;**3**:446.
10. Schlatter R, Schmich K, Avalos Vizcarra I, *et al*. ON/OFF and beyond–a boolean model of apoptosis. *PLoS Comput Biol* 2009;**5**(12):e1000595.
11. Ríos O, Frias S, Rodríguez A, *et al*. A Boolean network model of human gonadal sex determination. *Theor Biol Med Model* 2015;**12**:26.
12. Martinez-Sanchez ME, Mendoza L, Villarreal C, *et al*. A minimal regulatory network of extrinsic and intrinsic factors recovers observed patterns of CD4+ T cell differentiation and plasticity. *PLoS Comput Biol* 2015;**11**(6):e1004324.
13. Cohen DP, Martignetti L, Robine S, *et al*. Mathematical modelling of molecular pathways enabling tumour cell invasion and migration. *PLoS Comput Biol* 2015;**11**(11):e1004571.
14. Grieco L, Calzone L, Bernard-Pierrot I, *et al*. Integrative modelling of the influence of MAPK network on cancer cell fate decision. *PLoS Comput Biol* 2013;**9**(10):e1003286.
15. Remy E, Rebouissou S, Chaouiya C, *et al*. A modeling approach to explain mutually exclusive and co-occurring genetic alterations in bladder tumorigenesis. *Cancer Res* 2015;**75**(19):4042–52.
16. Steinway SN, Zañudo JGT, Ding W, *et al*. Network modeling of TGFβ signaling in hepatocellular carcinoma epithelial-to-mesenchymal transition reveals joint sonic hedgehog and Wnt pathway activation. *Cancer Res* 2014;**74**(21):5963–77.
17. Albert R, Thakar J. Boolean modeling: a logic-based dynamic approach for understanding signaling and regulatory networks and for making useful predictions. *Wiley Interdiscip Rev Syst Biol Med* 2014;**6**:353–69.
18. Abou-Jaoudé W, Monteiro PT, Naldi A, *et al*. Model checking to assess T-helper cell plasticity. *Front Bioeng Biotechnol* 2014;**2**:86.
19. Mbodj A, Junion G, Brun C, *et al*. Logical modelling of Drosophila signalling pathways. *Mol Biosyst* 2013;**9**(9):2248.
20. Chaouiya C, Bérenguier D, Keating SM, *et al*. SBML qualitative models: a model representation format and infrastructure to foster interactions between qualitative modelling formalisms and tools. *BMC Syst Biol* 2013;**7**:135.
21. Flobak Å, Baudot A, Remy E, *et al*. Discovery of drug synergies in gastric cancer cells predicted by logical modeling. *PLOS Comput Biol* 2015;**11**(8):e1004426.
22. Shannon P, Markiel A, Ozier O, *et al*. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;**13**:2498–504.
23. Chaouiya C, Naldi A, Thieffry D. Logical modelling of gene regulatory networks with ginsim. In: *Bacterial Molecular Networks*. New York, NY: Springer, 2012, 463–79.
24. Funahashi A, Morohashi M, Kitano H, *et al*. CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. *Biosilico* 2003;**1**(5):159–62.
25. Perfetto L, Briganti L, Calderone A, *et al*. SIGNOR: a database of causal relationships between biological entities. *Nucleic Acids Res* 2016;**44**(D1):D548–54.
26. Türei D, Korcsmáros T, Saez-Rodriguez J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat Methods* 2016;**13**:966–7.
27. Terfve CD, Wilkes EH, Casado P, *et al*. Large-scale models of signal propagation in human cells derived from discovery phosphoproteomic data. *Nat Commun* 2015;**6**:8033.
28. Stoll G, Caron B, Viara E, *et al*. MaBoSS 2.0: an environment for stochastic Boolean modeling. *Bioinformatics* 2017;**33**(14):2226–8.
29. Müssel C, Hopfensitz M, Kestler HA. BoolNet—an R package for generation, reconstruction and analysis of Boolean networks. *Bioinformatics* 2010;**26**(10):1378–80.
30. Terfve C, Cokelaer T, Henriques D, *et al*. CellNOptR: a flexible toolkit to train protein signaling networks to data using multiple logic formalisms. *BMC Syst Biol* 2012;**6**:133.
31. R Core Team. *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2014.
32. Batt G, Besson B, Ciron PE, *et al*. Genetic network analyzer: a tool for the qualitative modeling and simulation of bacterial regulatory networks. In: *Bacterial Molecular Networks*. New York, NY: Springer, 2012, 439–62.
33. Klarner H, Streck A, Siebert H. PyBoolNet: a python package for the generation, analysis and visualization of Boolean networks. *Bioinformatics* 2017;**33**:770–72.
34. Di Cara A, Garg A, De Micheli G, *et al*. Dynamic simulation of regulatory networks using SQUAD. *BMC Bioinformatics* 2007;**8**:462.
35. Mendes ND, Monteiro PT, Carneiro J, *et al*. Quantification of reachable attractors in asynchronous discrete dynamics. *ArXiv* 2014;abs/1411.3539:19. http://arxiv.org/abs/1411.3539.
36. Bonnet E, Calzone L, Rovera D, *et al*. BiNoM 2.0, a cytoscape plugin for accessing and analyzing pathways using standard systems biology formats. *BMC Syst Biol* 2013;**7**:18.
37. Gorban AN, Pitenko A, Zinovyev A. ViDaExpert: user-friendly tool for nonlinear visualization and analysis of multidimensional vectorial data. *ArXiv* 2014;abs/1406.5550:9. http://arxiv.org/abs/1406.5550.
38. Martignetti L, Calzone L, Bonnet E, *et al*. ROMA: representation and quantification of module activity from target expression data. *Front Genet* 2016;**7**:18.

39. Bonnet E, Viara E, Kuperstein I, *et al*. NaviCell web service for network-based data visualization. *Nucleic Acids Res* 2015; **43**(W1):W560–5.
40. Subramanian A, Tamayo P, Mootha VK, *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 2005; **102**:15545–50.
41. De Smet R, Marchal K. Advantages and limitations of current network inference methods. *Nat Rev Microbiol* 2010;**8**(10): 717–29.
42. Bonnet E, Calzone L, Michoel T, Gardner PP. Integrative multi-omics module network inference with Lemon-Tree. *PLoS Comput Biol* 2015;**11**(2):e1003983.
43. Faith JJ, Hayete B, Thaden JT, *et al*. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol* 2007;**5**(1):e8.
44. Chen X, You ZH, Yan GY, *et al*. IRWRLDA: improved random walk with restart for lncRNA-disease association prediction. *Oncotarget* 2016;**7**(36):57919–31.
45. You ZH, Huang ZA, Zhu Z, *et al*. PBMDA: a novel and effective path-based computational model for miRNA-disease association prediction. *PLoS Comput Biol* 2017;**13**(3):e1005455.
46. Weinstein N, Mendoza L. Building qualitative models of plant regulatory networks with SQUAD. *Front Plant Sci* 2012;**3**:72.
47. Morris MK, Melas I, Saez-Rodriguez J. Construction of cell type-specific logic models of signaling networks using CellNOpt. In: *Computational Toxicology*. Totowa, NJ: Humana Press, 2013, 179–214
48. Stoll G, Viara E, Barillot E, *et al*. Continuous time Boolean modeling for biological signaling: application of Gillespie algorithm. *BMC Syst Biol* 2012;**6**:116.
49. Abou-Jaoudé W, Traynard P, Monteiro PT, *et al*. Logical modeling and dynamical analysis of cellular networks. *Front Genet* 2016;**7**:94.
50. Lé S, Josse J, Husson F. FactoMineR: an R package for multivariate analysis. *J Stat Softw* 2008;**25**:1–18.
51. Zinovyev A, Viara E, Calzone L, *et al*. BiNoM: a cytoscape plugin for manipulating and analyzing biological networks. *Bioinforma Oxf Engl* 2008;**24**:876–7.
52. Chanrion M, Kuperstein I, Barrière C, *et al*. Concomitant Notch activation and p53 deletion trigger epithelial-to-mesenchymal transition and metastasis in mouse gut. *Nat Commun* 2014;**5**:5005.
53. Calzone L, Barillot E, Zinovyev A. Predicting genetic interactions from Boolean models of biological networks. *Integr Biol* 2015;**7**(8):921–9.
54. Foubert E, De Craene B, Berx G. Key signalling nodes in mammary gland development and cancer. The Snail1-Twist1 conspiracy in malignant breast cancer progression. *Breast Cancer Res* 2010;**12**:206.
55. Montagud A, Navarro E, Fernández de Córdoba P, *et al*. Reconstruction and analysis of genome-scale metabolic model of a photosynthetic bacterium. *BMC Syst Biol* 2010; **4**: 156.
56. Saez-Rodriguez J, Alexopoulos LG, Zhang M, *et al*. Comparing signaling networks between normal and transformed hepatocytes using discrete logical models. *Cancer Res* 2011;**71**(16): 5400–11.
57. Kanehisa M, Goto S, Sato Y, *et al*. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* 2012;**40**(D1):D109–14.
58. Croft D, Mundo AF, Haw R, *et al*. The Reactome pathway knowledgebase. *Nucleic Acids Res* 2014;**42**(D1):D472–7.
59. Kuperstein I, Bonnet E, Nguyen HA, *et al*. Atlas of cancer signalling network: a systems biology resource for integrative analysis of cancer data with Google Maps. *Oncogenesis* 2015;**4**: e160.
60. Vahid S, Thaper D, Gibson KF, *et al*. Molecular chaperone Hsp27 regulates the Hippo tumor suppressor pathway in cancer. *Sci Rep* 2016;**6**:31842.
61. Ou CY, LaBonte MJ, Manegold PC, *et al*. A coactivator role of CARM1 in the dysregulation of β-Catenin activity in colorectal cancer cell growth and gene expression. *Mol Cancer Res* 2011; **9**:660–70.
62. Basso K, Margolin AA, Stolovitzky G, *et al*. Reverse engineering of regulatory networks in human B cells. *Nat Genet* 2005;**37**: 382–90.
63. Nicolle R, Radvanyi F, Elati M. CoRegNet: reconstruction and integrated analysis of co-regulatory networks. *Bioinformatics* 2015;**31**(18):3066–8.
64. di Bernardo D, Thompson MJ, Gardner TS, *et al*. Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nat Biotechnol* 2005;**23**: 377–83.
65. Bonnet E, Tatari M, Joshi A, *et al*. Module network inference from a cancer gene expression data set identifies microRNA regulated modules. *PLoS One* 2010;**5**(4):e10162.
66. Mordelet F, Vert JP. SIRENE: supervised inference of regulatory networks. *Bioinforma Oxf Engl* 2008;**24**(16):i76–82.
67. Vermeirssen V, Joshi A, Michoel T, *et al*. Transcription regulatory networks in *Caenorhabditis elegans* inferred through reverse-engineering of gene expression profiles constitute biological hypotheses for metazoan development. *Mol Biosyst* 2009;**5**:1817–30.
68. Bonneau R, Facciotti MT, Reiss DJ, *et al*. A predictive model for transcriptional control of physiology in a free living cell. *Cell* 2007;**131**(7):1354–65.
69. Ciofani M, Madar A, Galan C, *et al*. A validated regulatory network for Th17 cell specification. *Cell* 2012;**151**(2):289–303.
70. Klamt S, Gilles ED. Minimal cut sets in biochemical reaction networks. *Bioinforma Oxf Engl* 2004;**20**(2):226–34.
71. Vera-Licona P, Bonnet E, Barillot E, *et al*. OCSANA: optimal combinations of interventions from network analysis. *Bioinforma Oxf Engl* 2013;**29**:1571–3.
72. Paulevé L. Pint: a static analyzer for transient dynamics of qualitative networks with ipython interface. In: Feret J, Koeppl H (eds). *Computational Methods in Systems Biology, CMSB*, Vol. 10545, Lecture Notes in Computer Science. Springer, Cham, 2017.