A logical approach to identify Boolean Networks that model cell differentiation

Stéphanie Chevalier, Andrei Zinovyev, Christine Froidevaux, Loïc Paulevé

Context

The gene interactions can be represented by networks, where gene activation and inhibition are modeled. By predicting the result of existing perturbations in diseases, these computational models of gene regulation bring a promising framework to suggest new therapeutic targets. However, the formal definition of causal networks that model the growth of a studied cell line is a major challenge.

Among the several modeling paradigms for networks, this work focus on qualitative models of dynamics, especially Boolean networks, where the activity of genes is seen as simply ON or OFF. Boolean networks allow capturing essential dynamical features, such as steady states and differentiation processes, whereas requiring few parameters compared to quantitative models. Moreover, by being more abstract, Boolean networks allow both to address large-scale networks and to derive robust predictions, which depend very little on precise quantitative features.

The inference of Boolean networks has been addressed in the literature for steady-state data and time-series data, on cell lines subjects to perturbations. However, no method allows to take into account differentiation features. Hence we explore model inference for cell differentiation to provide a scalable framework for the systematic identification of differentiation models. Considering the nature of differentiation data influences the model checking, the first step was to formalize the various contexts of cell differentiation data, according to the stability of the measured cells and the measurements representativeness.

The systematic approach aims at reducing biases of over-fitting, by understanding which part of the networks are crucial to reproduce the data (should be common to all found models), and part of the networks which are non-identifiable given the data (high variability among the admissible models). This approach rules out any try-and-test approaches which would simply iterate over all possible models and independently test their validity. Instead, we rely on logic programming to express in a same abstract model both the possible networks and the constraints they have to satisfy to fit the data depending on the contexts previously defined. The logical characterization of the network inference starts as an extension of a prior work [2] on model inference from time series data using Answer-Set Programming (ASP) that enumerates the solutions of the logical program.

After a brief description of the Boolean networks, we introduce different levels of interpretation of the differentiation data and we describe ensuing formal constraints. We apply our method on a network that models the central nervous system differentiation.

Boolean Networks

A Boolean Network (BN) is a pair of two sets [4] :

- a finite set of components (genes),
- the corresponding set of Boolean functions.

A component at time t takes the value either 1 (expressed) or 0 (not expressed). A state contains the overall expression level of all components of the Boolean network at time step t.

The state of each gene can be updated synchronously or asynchronously during a transition, respecting the Boolean functions. In this work we consider the asynchronous Boolean network (ABN) : one component is updated at a time.

A consecutive sequence of states obtained by state transitions is called a trajectory. A reachable state from a state x of the BN is a state that belong to a trajectory from x. State transition in ABN is non-deterministic : from a state, it is possible that there are several directly reachable states. Ultimately a trajectory reaches an attractor, which is defined as follows : the set of reachable states from a state of the attractor is the set of states of the attractor itself. The attractor is called a fixed point if it is a single state, and cyclic attractor otherwise. In our context, the attractors correspond to the long-term behavior of the cells and thus represent the phenotypes. A basin of attraction of an attractor is the set of all states that reach this attractor.

Various contexts of cell differentiation data

The differentiation data are gene expression measurements on cell populations or single cells, that reach various phenotypes depending on activation or inhibition of genes. To interpret it, we have formally defined what differentiation is, and we have observed that it corresponds to different experimental contexts, which influence the constraints to identify the Boolean networks that reproduce the data :

- The first level of contexts refers to hypothesis on the cell stability : it distinguishes whether measured cells were in a stable state ("steady observation") or not ("non-steady observation").
- The second level of contexts refers to hypothesis on the measurement representativeness : it distinguishes whether the sample is representative of the whole diversity of the population ("universal observation") or not ("existential observation").
- The third level of contexts refers to a particularity of the steady data : the measurement can correspond to a single state ("fixed point") or a cycle of states ("cyclic attractor"). And in this last specific case, another level of hypothesis can be made : the markers measured can be permanent or periodic in the attractor's states.

Depending on the context, the interpretation in constraints for the Boolean networks is different.

From interpretation to formal constraints

To infer qualitative models of differentiation mathematically expressed as Boolean networks, the first strategy would be to check the reachable states from the differentiated states. But checking the reachability of a state from another is a standard model-checking task, known to have a limited scalability due to its theoretical complexity (PSPACE-complete[1]) [2]. Hence our first concern is to translate data into constraints that have to be satisfied by the dynamics of a Boolean network model in order to be considered as a valid candidate model. Constraints depend on the contexts of the differentiation, and have to be expressed using logical programming with the aim of filter out quickly some true negatives (thanks to necessary conditions) and true positives (thanks to sufficient conditions), leaving the computational demanding validation to a subset of putative models.

Our current work focuses on the implementation of sufficient conditions. For instance let us consider a sufficient condition in the non-steady context, which consists of an over-approximation of the reachable states : if the over-approximation from the states that are differentiated is disjoined, we can already confirm that the Boolean network reproduces the differentiation, without a costlier check. The strategy actually slightly differs according to the hypothesis "universal/existential data", but the main current challenge is to translate it into logic programming, to extend the tool Caspots[2] that already has constraints on reachability.

The context of steady cells means in the Boolean network context that the observed states are in attractors. For the case of fixed point, an option has been added to Caspots, that allows to indicate time points as fixed points. For the case of cyclic attractor, constraint can be expressed in temporal logic (CTL) that allows a quite efficient network check.

Applications

To train our method we produce simulated data adapted to the manually-curated minimal network for central nervous system differentiation [3]. This network mainly consists of two steps of two mutually-inhibited fate decision genes : the first step determines the specialization in neuron or glia where the second step determines that of astrocyte and oligodendrocyte.

Infer on this 5-node network by considering time series data highlights 256 compatible Boolean networks. When we specify the specialization in neuron, glia and astrocyte as 3 fixed points thanks to our new option integrated in caspots, this constraint brings out that only 4 of these Boolean networks are actually compatible with the data.

Thus, when it is known that time points of time series data correspond to steady states, the fixed point constraint is relevant to discriminate efficiently Boolean networks.

Perspectives

The short-term focus is to find other constraints to implement in ASP in the differentiation contexts described, so as to infer network efficiently, even under realistic conditions with large gene networks. But other axes complete this work.

A first topic is that, up to now, we have only considered the existence of differentiation processes. We must also cover the experimental situation where differentiation leads to a change in the probabilities of reaching phenotypes.

Then in a longer-term view, being able to identify networks that reproduce the observed differentiation on time series data is a first step in the goal of network inference about CRISPR-Cas9 mutation data, for which the only dynamic information could be the phenotype of the cell population (proliferation, apoptosis, ...).

Références

[1] Allan Cheng, Javier Esparza, and Jens Palsberg. Complexity results for 1-safe nets. *Theoretical Computer Science*, 147(1) :117 – 136, 1995.

[2] Max Ostrowski, Loïc Paulevé, Torsten Schaub, Anne Siegel, and Carito Guziolowski. Boolean Network Identification from Perturbation Time Series Data combining Dynamics Abstraction and Logic Programming. *BioSystems*, 2016.

[3] Xiaojie Qiu, Shanshan Ding, and Tieliu Shi. From understanding the development landscape of the canonical fate-switch pair to constructing a dynamic landscape for two-step neural differentiation. *PLOS ONE*, 7(12) :1–14, 12 2012.

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