Formalizing and enriching phenotype signatures using Boolean networks

Méline Wery, Olivier Dameron, Jacques Nicolas, Elisabeth Rémy, Anne Siegel

Formalizing and Enriching Phenotype Signatures using Boolean Networks

Méline Wery[a,b], Olivier Dameron[a], Jacques Nicolas[a], Élisabeth Remy[c], Anne Siegel[a,*]

[a] *Univ Rennes, Inria, CNRS, IRISA F-35000 Rennes, France*
[b] *SANOFI R&D, Translational Sciences, Chilly Mazarin, 91385, France*
[c] *Aix Marseille Univ, CNRS, Centrale Marseille, I2M, Marseille, France*

Abstract

In order to predict the behavior of a biological system, one common approach is to perform a simulation on a dynamic model. Boolean networks allow to analyze the qualitative aspects of the model by identifying its steady states and attractors. Each of them, when possible, is associated with a phenotype which conveys a biological interpretation. Phenotypes are characterized by their signatures, provided by domain experts. The number of steady states tends to increase with the network size and the number of simulation conditions, which makes the biological interpretation difficult. As a first step, we explore the use of Formal Concept Analysis as a symbolic bi-clustering technics to classify and sort the steady states of a Boolean network according to biological signatures based on the hierarchy of the roles the network components play in the phenotypes. FCA generates a lattice structure describing the dependencies between proteins in the signature and steady-states of the Boolean network. We use this lattice (i) to enrich the biological signatures according to the dependencies carried by the network dynamics, (ii) to identify variants to the phenotypes and (iii) to characterize hybrid phenotypes. We applied our approach on a T helper lymphocyte (Th) differentiation network with a set of signatures corresponding to the sub-types of Th. Our method generated the same classification as a manual analysis performed by experts in the field, and was also able to work under extended simulation conditions. This led to the identification and prediction of a new hybrid sub-type later confirmed by the literature.

Keywords: Phenotype Signature, Dynamic Models, Formal Concept Analysis, Boolean networks, Steady state.

*Corresponding author

1. Introduction

Systems biology aims to understand how the interactions between cellular components determine the cell response to environmental perturbation by external stimuli. Historically, two main approaches have been developed to take into account the dynamical behavior of a regulatory network. Inspired by modeling technics used in physics, continuous models based on differential equations are widely used to investigate the role of circuits in regulatory and signaling networks, as well as the fluctuations of concentrations over time. Notice however that model calibrations are difficult and require a large set of quantitative experimental measurements which are hardly available in the context of regulatory interactions [23, 24, 7]. Another weakness of continuous and stochastic modeling technics is that they implicitly assume that reactions follow mass action kinetics, although regulatory interactions have been observed and measured to be rather similar to switches, or at least very sharp in terms of responses [35, 8].

To overcome these limitations, discrete frameworks have been introduced to describe the response of regulatory networks. A first class of synchronous models associate the gene or protein states to binary variables whose values are controlled by the binary values of their regulators [16]. This approximation of regulation appeared to be overly simplistic because it does not take into account the possibly different time-scales occurring in regulatory networks, nor the fact that a component might act differently according to its level of expression or activation. Multi-level and asynchronous logical formalisms, the so-called Thomas models [30, 31], have been proved to be accurate for modeling regulatory systems because they capture the main features of the different time-scales in regulatory processes with asynchronous and non-deterministic formalisms [23, 2].

An output of the study of a logical network by multi-level and asynchronous logical formalisms is the enumeration of its steady states and more generally the study of its attractors [14]. This feature has been widely used to link models of regulations with phenotypes, especially in health-related applications. In [10], it was proved that the attractors of a mammal cell-cycle in several perturbed conditions were in agreement with known phenotypes in the literature. The steady states and attractors of a logical model were also proved to fit with genotyping information in [25]. Finally, in [22, 1], the authors studied a network of T-helper lymphocytes and evidenced that the steady states of the network in several environmental or gene-deletion/activation conditions were in agreement with observed clinical phenotypes. These phenotypes were either generic (proliferation, apoptosis...) or were more specific and described subtle differences in cancer cell-types.

All these studies establish a link between some phenotypes (especially in cancer situations) and the steady states of a logical network. In concrete terms, this link provides a signature for each phenotype, that is, a set of biological markers present in the steady states whose activation is characteristic of the phenotype.

Notice however that the concept of signature is loosely defined in the literature depending on the context. A phenotype signature is generally defined as

the set of master genes or proteins characterizing this phenotype. Signatures can be computed according to gene set enrichment analysis and gathered in databases such as SigDB [18]. They can also be computed to focus on causality effects if logical networks are available [27, 17, 11]. However, most of the time, signatures are refined manually by clinicians or biological experts in order to be more accurate and discriminate cell-types according to a few biomarkers. This is for example the case of CD4 T helper cells (Th cells) for which several cell-types have been identified (Th1, Th2, Th17, Th9, Th22). The heterogeneity of Th cells is closely related to signals from the microenvironment; typically IL-12 is required for the development of Th1 cells; IL-4/IL-2 drive the development of Th2 subtype, and TGFb induces the differentiation of Th17 cells as well as T regulatory (Treg) cells. Moreover those main Th cells subtypes are associated with very specific biomarkers. Indeed Th1, Th2, Th17 are characterized by the expression of T-bet, Gata3 and RORgt respectively; while Treg cells are characterized by Foxp3 expression. Yet, Th cells present a certain degree of plasticity, and notably they can adopt hybrid phenotypes.

Hybrid phenotypes can appear when the biomarkers of several signatures are measured simultaneously in the same cell-type [9]. For instance the Th1-Th17 subtype expresses both T-bet and RORgt master regulators and produces both INFg and IL-17. Therefore, links between signatures, phenotypes and steady states in logical models become intricate as soon as the network size increases. In this situation, classification methods such as hierarchical clustering [36] highlight the main links between signatures and phenotypes, but fail to describe all the possible variants and hybrids that co-exist with the main clusters.

Our study aims at developing an automatic method to classify the steady states of a logical network according to a given family of phenotypes. These phenotypes are defined by their own signatures that can be either a single master gene or more generally a pattern of activated biological compounds (genes, proteins or markers). Our framework allows a systematic exploration of the links between phenotypes signatures and values of genes in the steady states of a Boolean network.

It relies on a classification of the steady states using a hierarchical structure derived from Formal Concept Analysis (FCA), a data analysis method handling binary matrices [13]. In our case, we focus on the analysis of matrices describing the list of activated compounds in steady states. The FCA method produces a lattice, representing in a hierarchical way associations as bi-clusters of specific nature, named *concepts*. Each concept is made of a subset of rows (in our cases, steady-states) which exhibit similar characteristics across a subset of columns (in our case, the compounds shared by the steady-states). By studying concepts associated with signatures of phenotypes, thanks to this hierarchical structure, we can perform a systematic characterization of biological compounds which are always paired with the signature's master genes or proteins according to the Boolean network. Hybrid phenotypes can also be characterized in association with all their possible variants. This is illustrated on several models representing the differentiation of LTh cells [26, 22, 1]. Interestingly, the classification enables the identification of novel hybrid types together with the simulation conditions

that generated them.

2. Organizing the steady states of a Boolean network into a lattice

2.1. Network representation and simulation

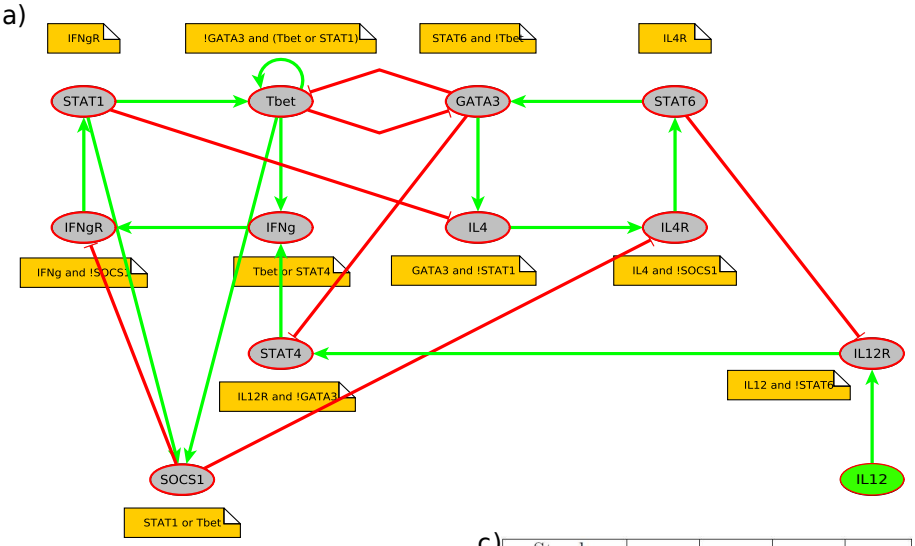
R. Thomas has proposed a logical formalism [30, 31] to model regulatory networks. It is based on two directed graphs and a system of logical rules coding for the network dynamics. The interaction network is represented by a *regulatory graph* (RG), whose nodes stand for the biological compounds of the system, and edges stand for the interactions between these components (transcriptional activation or inhibition). In this work, we distinguish two types of nodes, external nodes (also called input nodes) and internal nodes. External nodes represent the *input* of the conditions of simulation. These compounds can not be regulated during the dynamics but their values can be fixed. To each internal compound are attached (1) a discrete variable representing the expression of the biological compound qualitatively (its *state*). We consider here only Boolean variables; (2) a logical function depicting the evolution of the component with respect to the states of its regulators. If the values of the internal compounds are specified at the beginning of the simulation, this state is called an *initial state*. The *State Transition Graph* (STG) represents the discrete dynamics; nodes are the states of the system, and transitions link two consecutive states [30]. Hence, the STG encompasses all the possible trajectories with respect to the set of logical functions parametrizing the RG under the asynchronous hypothesis (*i.e.* only one component differs between two consecutive states). *Attractors -i.e.* terminal strongly connected components of the STG- are parts of the STG where the system stabilizes, interpreted as the long-term behavior of the system. There are two types of attractors: *steady states* and *cyclic attractors*, according to whether they are made up of one or several states.

Once the regulatory graph and its logical rules are defined, it is possible to run simulations specifying (or not) initial conditions using GINsim [14]. This software offers several functions such as the computation of all steady states, the reduction of the model based on a compressed STG, and various simulations according to pre-specified mutations. We have implemented a Python script that generates and extracts all the steady states of a model encoded in GINsim. Steady states are represented as rows of a Boolean matrix, whose columns correspond to the genes/proteins. The (i, j)-th coefficient of the matrix is equal to 1 if and only if the component j is expressed in the steady state i. An example is shown in Fig. 1.

2.2. Formal Concept Analysis (FCA)

Our goal is to compare the steady states of a logical network over multiple simulations with different initial conditions. Moreover we are interested in an exhaustive enumeration of possible clusters of states without *a priori*. This involves, in particular, managing cluster overlaps.

a)



b)

Simulation 0	None (IL12 = 0)
Simulation 1	IL12 (IL12 = 1)

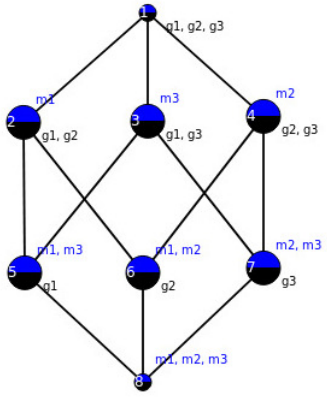
c)

Steady States	s_1	s_2	s_3	s_4
Simulation Condition	None	None IL12	None	IL12
IFNγ	0	0	1	1
IL4	0	1	0	0
IFNγR	0	0	0	0
IL4R	0	1	0	0
IL12R	0	0	0	1
STAT1	0	0	0	0
STAT6	0	1	0	0
STAT4	0	0	0	1
SOCS1	0	0	1	1
GATA3	0	1	0	0
Tbet	0	0	1	1

Figure 1: Small-scale network controlling the differentiation of Lymphocyte T helper (Th) with two input environments (data extracted from [26]). (a) The network (has 11 internal compounds (gray nodes) and one external compound IL12 (green node)). Yellow labels give the logical function of each compound. Activations are represented by green arrows and inhibition by red arrows. (b) Input conditions used during simulation. The two values of the external node or input IL12 correspond to the stimulation or not of this gene.(c) Matrix crossing compounds and steady states with the conditions of stimulation (value 1 for an activated compound in a state, 0 otherwise): the first condition of stimulation (IL12=0) generated the three steady-states s_1, s_2 and s_3 whereas s_2 and s_4 are the two steady-states which can be accessed with the condition IL12=1. In both cases, the convergence of the dynamics to one or the other steady-states depends on the initial state of the internal nodes.

	m_1	m_2	m_3
g_1	×		×
g_2	×	×	
g_3		×	×

(a) Formal context



(b) Concept lattice

Figure 2: Example of a concept lattice. Table (a) describes a relation between the set of objects $G = \{g_1, g_2, g_3\}$ and the set of attributes $M = \{m_1, m_2, m_3\}$. Concepts are nodes in the graph (b) labelled with subsets of G and M, drawn with *LatViz* [3]. For instance, the top left is $(\{g_1, g_2\}, \{m_1\})$. Indeed, both g_1 and g_2 are in relation with m_1, and it is not possible to add g_3 ($g_3 \times m_1$ is lacking), m_2 ($g_1 \times m_2$ is lacking) or m_3 ($g_2 \times m_3$ is lacking). Similarly, the seven other formal concepts are $(\{g_1, g_2, g_3\}, \{\})$ -top-, $(\{g_2, g_3\}, \{m_2\})$, $(\{g_1, g_3\}, \{m_3\})$, $(\{g_2\}, \{m_1, m_2\})$, $(\{g_1\}, \{m_1, m_3\})$, $(\{g_3\}, \{m_2, m_3\})$ and $(\{\}, \{m_1, m_2, m_3\})$ -bottom-. They are organized in a lattice according to the set inclusion relationship between objects.

Formal Concept Analysis (FCA) is a widely used data analysis technics that can be used for this purpose. In its most simple form, concepts formalize the duality extension and intension by extracting, from a binary relation between a set of objects and a set of attributes, the maximal subsets of objects that share the same subset of attributes [12]. Causality relations can be investigated within a lattice structure (Galois connection) by subconcept-superconcept relations. In bioinformatics, it has been used to derive phylogenetic relations among groups of organisms [19] and to exhibit clusters in large-scale interaction networks [5, 34]. In the following, objects are steady states of a Boolean network, and attributes are activations of biological compounds.

Formally, a data table allows building a context (G, M, I) where G (objects) and M (attributes) are two finite sets and $I \subset G \times M$ describes a relation between G and M. The set of attributes shared by all elements of $A \subset G$ is denoted by $A' = \{m \subset M | A \times \{m\} \subset I\}$. Similarly, the set of objects sharing all the elements of B is denoted by $B' = \{g \in G | \{g\} \times B \subset I\}$. The pair (A, B) is called a *formal concept* if $B = A'$ and $A = B'$, A being the extent and B the intent of the concept. Additionally, the extent and the intent are closed sets, i:e A = A" and B = B". Equivalently and more intuitively, (A, B) is a formal concept precisely when every object in A is in relation with every attribute in B and it is not possible to add an element to A or B without breaking this property. For instance, for the relation shown in Fig.2, the concept lattice is

a Boolean lattice with 2^3 nodes. The formal concept associated with m_1, is $(\{g_1, g_2\}, \{m_1\})$ and none of the elements g_3, m_2 and m_3 can be added to the concept. In this work, formal concepts identify all sets of states that share the same biological elements.

For a set of objects $A \subset G$, we associated to A an *attribute concept*, that is, the smallest formal concept which contains A, that is to be, (A'', A'). Dually, for a set of attributes $B \subset G$, the *object concept* is the largest formal concept which contains B, that is to be (B', B''). This allows us to identify either all the steady states sharing the same features, or all the biological elements characterizing a set of steady states. For the relation shown in Fig.2, there are six non-trivial formal concepts.

A partial order is defined over the family of formal concepts by $(A_1, B_1) \leq (A_2, B_2) \iff A_1 \subset A_2$ (or equivalently $B_2 \subset B_1$). The set of concepts forms a lattice. It means that every pair $((A_1, B_1), (A_2, B_2))$ of formal concepts has a greatest common subconcept $((A_1 \cap A_2), (B_1 \cup B_2)'')$, the *meet* and a lowest common superconcept $((A_1 \cup A_2)'', (B_1 \cap B_2))$, the *join*. The lattice can be represented as a graph. As shown in Fig.2b, formal concept are represented by a node in this graph. The top node is the concept with all objects and the bottom node is the concept with all attributes.

The lattice provides information about the impact of the addition or the deletion of objects or attributes over the classification process. More precisely, the structure of the concept lattice is suitable for capturing causalities through the use of implications and associations rules extracted from the lattice [13]. Without entering into details, we will implicitly use these causalities in the following section to explore the lattice associated with a Boolean network.

2.3. Building a lattice from a family of states in a Boolean Network

In the following, the term "compound" encompasses a gene, a protein or a biological compound. All are nodes in the regulatory network. In the following, the term "node" will refer to a formal concept in the lattice.

Let ϕ be a Boolean Network (BN) over a set of compounds $\boldsymbol{V}$ with values in $\{0, 1\}$. Let $\boldsymbol{S} \subset \{0, 1\}^{|\boldsymbol{V}|}$ be a family of steady states of ϕ (in general, it could be any subset of states). We define the relation I on $\boldsymbol{S} \times \boldsymbol{V}$ by setting that $(s, v) \in I$ if and only if v is activated in steady state s.

From context $(\boldsymbol{S}, \boldsymbol{V}, I)$ we build a lattice of formal concepts. As an example, Fig. 3 shows the formal context associated with 10 states of a BN. As shown in Fig. 3(b), a hierarchical clustering approach applied to the considered table highlights the role of the cluster $\{s_2, s_3, s_5, s_6, s_7, s_9\}$. A follow-up manual analysis suggests that the system's states are the states for which the compounds $\{b, c, g, i, l\}$ are simultaneously activated. On the contrary, the associated formal concept lattice in Fig.3(c) provides an exhaustive view of all existing relations between states and compounds. It contains 26 concepts corresponding to 26 subsets of objects which exhibit similar characteristics across their associated subset of attributes. This enables the reconstruction of associations. For instance, the red concept associated with $\{d\}$ contains states

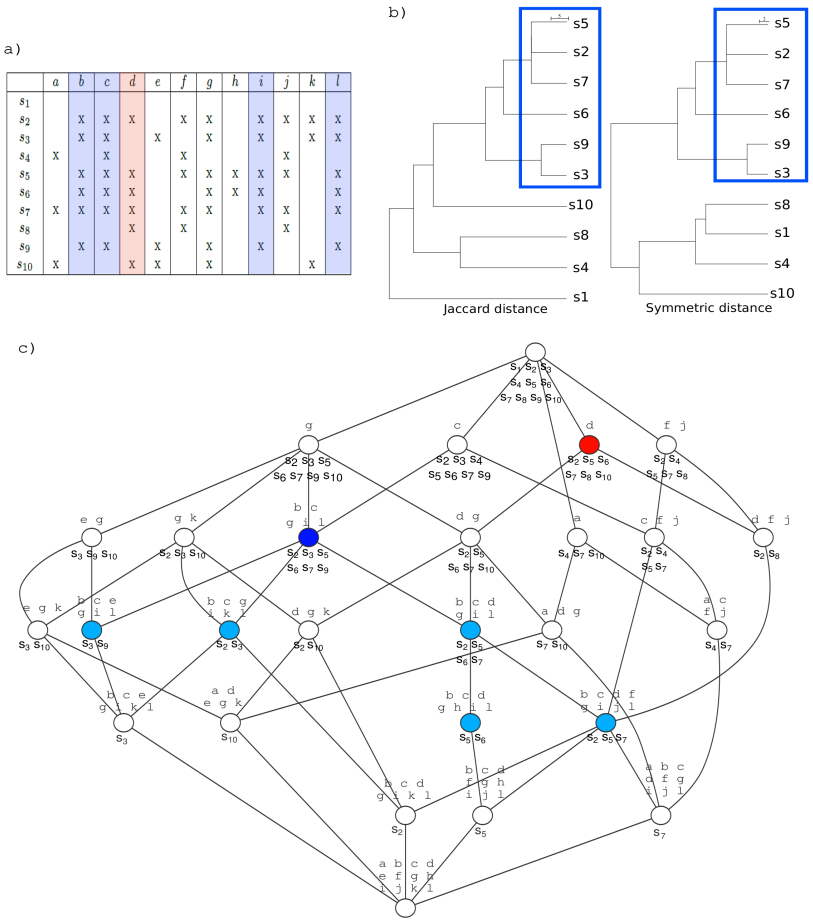


Figure 3: 10-states matrix of a Boolean network and comparison between FCA and UGPMA clustering (a) steady states are structured into a matrix where columns are compounds and rows states of the system. A compound is activated in a state if the corresponding cell in the matrix is checked (×). (b) Hierarchical clustering of states based on the Jaccard and symmetric distances. The blue frame characterizes the main common cluster, associated with the signature $\{b, c, g, i, l\}$. (c) Concept lattice derived from the matrix. The concept associated with $\{d\}$ is the red concept with objects $\{s_2, s_5, s_6, s_7, s_8, s_{10}\}$. These are the states for which the compound d is activated according to the table. The dark blue concept is associated with the signature $\{b, c, g, i, l\}$. It describes the family of six states for which the 5 compounds in the signature are activated. We notice that no formal concept contains the family of 4 compounds $\{b, c, i, l\}$. This means that the fifth compound g is always activated whenever the four compounds b, c, i, l are activated. and it should be added to the signature. There are five light blue concepts below the formal concept associated with $\{b, c, g, i, l\}$. They are called *variants* and can be automatically identified using the lattice structure.

$\{s_2, s_5, s_6, s_7, s_8, s_{10}\}$: they are the states for which d is activated according to the table. Similarily, the lattice shows that the formal concept associated with $\{b\}$ (dark blue node in the lattice) has $\{b, i, l, c, g\}$ as set of attributes, which corresponds to the compounds identified above by post-processing the result of the clustering methods. Because there does not exist a formal concept (or node) with only $\{i\}$ or $\{l\}$, this means that b, i and l cannot be distinguished according to the relation provided by the table: they have similar columns. In the FCA formalism, we have there attribute equivalences (b is equivalent to i which is equivalent to l) and implication rules $b, g, i, l \implies c$ and $b, c, i, l \implies g$. Such an information, which is highly valuable for the study of dependencies in a network, cannot be obtained with clustering approaches and will be exploited all along our study.

As shown in this example, the main advantage of the lattice structure is to gather an exhaustive representation of the state families, and their inclusion relations, according to the activated compounds they have in common: instead of using a strong statistical signal it rather explores all dependencies between states and compounds, allowing these dependencies to be propagated along the lattice in order to investigate the role of the deletion or the addition of an activated compound in the clustering process.

3. Exploring the lattice of steady states according to biological signatures of phenotypes

3.1. Refinement of signatures according to phenotype knowledge

As stated in the introduction, it is common to interpret the different attractors of a model through one or several signatures, that is, sets of proteins or genes whose simultaneous activation is interpreted as a characteristic of a particular phenotype. In the following, we will denote by $\mathcal{S}g = \{v_1, ..., v_n\} \subset \boldsymbol{V}$ a phenotype signature, possibly provided by an expert.

A first added-value of FCA is to allow a systematic interpretation of the steady states with respect to a signature $\mathcal{S}g$. As the signature $\mathcal{S}g$ is a set of compounds, it can be associated to a unique nearest concept in the lattice, that is, the formal concept whose set of attributes contains all the biological compounds of $\mathcal{S}g$ and is minimal with respect to this property. As an example, let us assume that a phenotype is characterized by signature $\{b, c, i, l\}$. The lattice depicted in Fig.3(c) highlights that the greatest formal concept which contains these compounds as attributes is $\{s_2, s_3, s_5, s_6, s_7, s_9\} \times \{b, c, g, i, l\}$. This concept conveys two informations. First, it points out to the six states that satisfy the signature. According to the notations introduced in Sec. 2.2, these states correspond to $\{b, c, i, l\}' = \{b, c, g, i, l\}'$. Second, it states that the activation of the compound g occurs whenever the 4 biological compounds in the signature are activated. These compounds correspond to $\{b, c, i, l\}$". As a matter of interpretation, this suggests that a refined signature, with the family of studied phenotypes (i.e., states), is $\{b, c, g, i, l\}$.

Based on this example, we define a refined signature of $\mathcal{S}g$ according to the set of steady states $\boldsymbol{s}$ to be the minimal set of attributes of the formal

concept associated with $\mathcal{S}g$ in the lattice. With this definition, $\{b, c, g, i, l\}$ is the refined signature of any sub-family of $\{b, c, g, i, l\}$ which does not appear itself in the lattice or in another super-concept of the refined signature depicted in Fig.3(c). On the contrary, the refined signature of $\{c\}$ is $\{c\}$ itself, since the lattice contains the concept $\{s_2, s_3, s_4, s_5, s_6, s_7, s_9\} \times \{c\}$.

3.2. Variants

As explained above, the signature of a phenotype is the set of master genes or proteins characterizing this phenotype. Clearly, it may appear that several cells share the same master genes -and thus have the same canonical phenotype- although they differ by other "minor" (i.e. not master) components. Thus, the set of cells associated with a phenotype may contain subclasses, characterized by a subset of minor components. They constitute variants of the same canonical class. We extend the notion of refined signature to variants: we formally define the variants of a signature to be sets of attributes associated with concepts that are smaller than the concept associated with the refined signature and contain at least two states (to avoid signatures specific of a single state).

In the example depicted in Fig.3(c) for instance, starting from the biological signature $\{b, c, i, l\}$, we obtained the refined signature $\{b, c, g, i, l\}$, and this signature has five variants. The formal concept $\{b, c, e, g, i, l\} \times \{s_3, s_9\}$ allows stating that $\{b, c, e, g, i, l\}$ is the variant signature corresponding to the activated compounds shared by the states s_3 and s_9. Other variants are provided by concepts $\{b, c, g, i, k, l\} \times \{s_2, s_3\}$, $\{b, c, d, g, i, l\} \times \{s_2, s_5, s_6, s_7\}$, $\{b, c, d, g, h, i, l\} \times \{s_5, s_6\}$ and $\{b, c, d, f, g, i, j, l\} \times \{s_2, s_5, s_7\}$. This example illustrates that variants defined by the FCA framework correspond to all combinations of minor components shared by several phenotypes which satisfy the signature, and they eventually facilitate the understanding of the role of these additional markers. In Fig.3(b), we computed the supervised classification obtained with UGPMA clustering based on two metrics, Jaccard and symmetric distance for the 10 considered states. As shown in this clustering, the main signature $\{b, c, e, g, i, l\}$ can be easily identified (blue frame) and is related to one specific node of those hierarchical trees. Variants are the sub-clusters contained in this blue frame depicted by the nodes. We notice that three variants are identified by both approaches : $\{b, c, e, g, i, l\} \times \{s_3, s_9\}$, $\{b, c, d, g, i, l\} \times \{s_2, s_5, s_6, s_7\}$, $\{b, c, d, f, g, i, j, l\} \times \{s_2, s_5, s_7\}$. They are all variants/clusters found in the UGPMA clustering. However, some variants like $\{b, c, g, i, k, l\} \times \{s_2, s_3\}$ and $\{b, c, d, g, h, i, l\} \times \{s_5, s_6\}$ can not be recovered in the hierarchical clustering. This illustrates the impact of the metric used in hierarchical clustering in terms of interpretation and classification. FCA uses a combinatorial approach rather than a statistical-based selection and performs a complete enumeration of possible variants for a given phenotype (or signature).

3.3. Identifying hybrids of several phenotypes characterized by their signatures

In this section, we will explain how FCA is suitable also for the analysis of sets of signatures. We assume that the steady states reached from different simulation conditions correspond to different canonical cell types, each associated

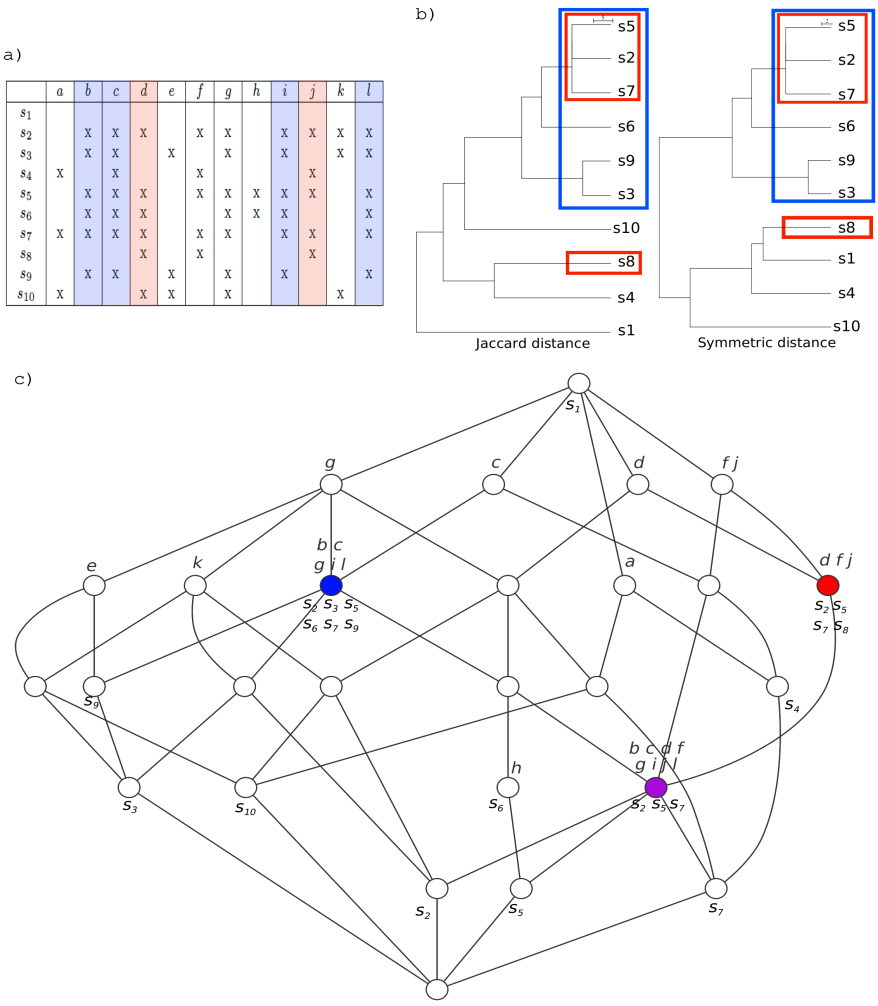


Figure 4: Hybrid of two cell-types characterized by their signatures (a) A Boolean matrix, the activated genes or proteins of the states of a Boolean network. A first phenotype is characterized by the biological signature $\{b, c, i, l\}$ (blue). Its refined signature is $\{b, c, g, i, l\}$, shared by states $\{s_2, s_3, s_5, s_6, s_7, s_9\}$. A second phenotype signature is characterized by $\{d, j\}$ (red). Its refined signature is $\{d, f, j\}$, shared by $\{s_2, s_5, s_7, s_8\}$. (b) Hierarchical clustering (average linkage) obtained from 2 distance matrices computed from the binary table (Jaccard and symmetric difference distance). The blue frame represents the set of states associated with the first signature and the red frame the set of steady states associated with the second signature. None of the metrics shows the link between s_8 and $\{s_2, s_5, s_7\}$. (c) Concept lattice associated with the matrix according to FCA. The concept associated with signature $\{b, c, i, l\}$ is in dark blue,the concept associated with the second signature $\{d, j\}$ is in red and the concept associated with the hybrid is in purple.

with a biological signature characterized by master genes or proteins. According to the definitions introduced in the previous sections, each biological signature can be extended to a contextually-refined signature and can be enumerated.

One should notice that the definition of the refined-signature and the variants depend on the simulation conditions of the network. It has been observed in several studies that by modifying inputs (environments) and/or initial conditions, one may obtain steady states with more than one master gene [22], and possibly hybrid steady states for which several master genes associated with different cell-types signatures are activated. This corresponds to the concept of hybrid cell types in the literature [28, 20]. For instance, [28] showed that a subpopulation of dendritic cells shared several surface markers with macrophages and appeared in the tumor microenvironment.

We define hybrid concepts to be concepts which are variants of at least two different cell-type signatures. We create a new hybrid cell type from two cell types t_1 and t_2 if and only if the meet $C_1 \wedge C_2$ of concepts C_1 and C_2, containing the refined signatures of t_1 and t_2, differs from the smallest concept in the lattice (bottom). The signature of the hybrid cell-type is the set of attributes of $C_1 \wedge C_2$. By construction, all genes and proteins in the signature belong to both t_1 and t_2 signatures. This concept itself may have variants, which are variants common to t_1 and t_2.

Fig.4 is an extension of the example shown in Fig.3. In this case, in addition to the cell-type defined by signature $\{b, c, i, l\}$, we consider another cell-type characterized by signature $\{d, j\}$. First, if we look at the UPGMA clustering [4], it appears that $\{d, j\}$ (red frame) is not identified by any of the metrics. Indeed, the steady state s_8 is isolated. However, the lattice shows that this second cell type has extended signature $\{d, f, j\}$ (red concept), with one variant, $\{b, c, d, f, g, i, j, l\}$ (purple concept). Importantly, this variant is also a variant of the first cell-type (blue concept). Therefore, both cell-types have an hybrid, with signature $\{b, c, d, f, g, i, j, l\}$. It differs from cell type $\{d, j\}$ by forcing the activation of $\{b, c, g, i, l\}$ and from cell type $\{b, c, i, l\}$ by forcing the activation of $\{d, f, j\}$. Interestingly, this cell-type has itself four variants which have no common compound with the hybrid and are called canonical variants to the cell-type $\{b, c, i, l\}$.

Together, the FCA framework allows the computation of all hybrid signatures for any family of signatures. Note that given a number of cell types n, the number of their variants can be exponential in n but the number of hybrids can be at most quadratic in n. In biological applications, hybrids are expected to be scarce and therefore may be easily analyzed manually.

3.4. Implementation

We implemented a python package *Foclass* to compute the concept lattice associated with a Boolean Network (available at `https://github.com/mwery/Foclass`). The Python package takes as input a Boolean Network together with simulation conditions formatted as a GINsim archive and a list of biological signatures provided as a text file. Those signatures are a set of activated genes associated with a biological phenotype as in [22]. The aim of the pipeline is to

automatically classify all the steady states generated from the BN, according to the signatures and using the FCA.

The *ginsimToInputFile* command computes steady states for a given Boolean network. To that goal, the GINsim software (command line interface) computes all the steady states and the script generates the resulting matrix crossing states and compounds (genes or proteins). Each cell contains a Boolean value stating the presence of a compound in a state.

The analysisFCA command performs the FCA on the matrix: steady states are objects, compounds are attributes, null values are empty cells. The list of concepts is computed with a dedicated Python package and can be used to classify steady states. In order to improve the performances of the algorithm, the computation focuses on the list of concepts and omits the lists of relations within the lattice. This command analyses the set of formal concepts according to the classification of signatures. The family of input signatures is extended based on the network phenotypes provided by the GINSim simulation by selecting concepts with the smallest set of compounds containing this signature. Concepts associated with hybrid phenotypes are then listed, allowing for the classification of all steady states which either satisfy a single canonical signature or belong to hybrid phenotypes. The full exploration of the set of formal concepts also enables computing the number of variants – formal concepts that contain, at least, the signature in their attributes – for canonical and hybrid signatures. When the numbers of objects is less or equal than 300, the lattice can be drawn to show canonical and hybrid signatures.

4. Application to Th cells differentiation - Exhaustive and automatic study of hybrid phenotypes

We evaluate our tool on the Th cell differentiation process. Indeed, this well studied differentiation process generates a large number of different canonical cell types characterized by a (set of) master genes, proteins, or markers which constitute the cell type signatures. The different phenotypes have been proved to be associated with several conditions of simulations of logical networks of various complexities. They therefore provide a relevant case-study to demonstrate the added-value of FCA-based analyses.

4.1. Biological context

T helper cells are lymphocytes that mature in the thymus and play a central role in the adaptive immune system. There are several subsets of Th cells; each type has been shown to express different cytokine profiles driving different immune response. Three main canonical cell types were first identified [21] : Th0 the naive form; Th1 a pro-inflammatory type which expresses the specific transcription factor T-bet and produces IFN-γ; and Th2, involved in allergic responses, which is induced by GATA3 expression and produces several interleukines (IL4, IL5). Over the last decade, several additional Th subtypes have been discovered : regulatory T cells (Treg), which depend on FOXP3 expression, and Th17 cells, induced by RORγT expression [32]. More recently, three

additional subsets have been characterized [15, 6]. Th9 is linked to PU.1 expression and can be differentiated both from Th2 with stimulation of TGFβ and from Th0 with combination of TGFβ and IL-4. Th22 can be induced by stimulation of Th0 with TNF-α and IL-6 which drive STAT3 expression. Finally, T follicular helper cells (TFH) depend on several cytokine stimulations which cause BCL-6 expression.

4.2. Identification of variants in a small case study

a)

Phenotypes	Associated signature
Th0	No protein = "empty"
Th1	Tbet
Th2	GATA3

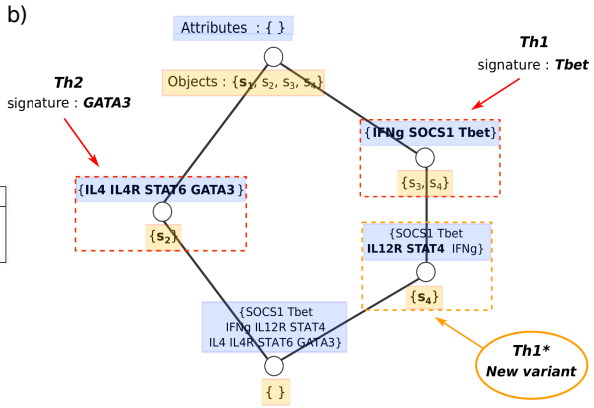


Figure 5: Concept lattice associated with the small-scale network controlling the differentiation of Lymphocyte T helper (Th). This network was described in Fig. 1 together with the four steady states reachable in two simulation conditions. (a) Definition of the signature for each cell type. Th1 and Th2 are characterized by the expression of a single master regulator. The signature of Th0 is empty since it reflects the absence of expressed protein. (b) Concept lattice generated from the matrix with compounds (attributes) in blue frame and states (objects) in orange frame. The two red hatched rectangles represent the two formal concepts associated with the signatures of Th1 and Th2. The orange hatched rectangle shows the formal concept associated with the variant of Th1.

Different models have been developed aiming at understanding the differentiation into each Th type under microenvironment change, and particularly in Boolean or multilevel framework [26, 22, 1]. As a first approach, [26] introduced a simplified model of the transcriptional regulatory network involved in the differentiation of $CD4^+$ naive T lymphocyte (Th) into Th1 or Th2, two active forms (Fig. 1(a)). Those two differentiated cell-types are triggered by two master genes (transcriptional factors): Tbet for Th1 and GATA3 for Th2 (Fig. 5(a)). This network includes 12 Boolean internal components and one input component, IL12, which represents the cellular environment and is known to induce the differentiation of Th0 into Th1 or Th2 (Fig. 1(a)). Dynamical simulations can be performed for each possible value of input (IL12=0 and IL12=1) (Fig. 1(b)). They lead to that four different steady states, shown in Fig. 1(c) and in the associated concept lattice in Fig. 5(e). According to the lattice and

the definitions introduced in the previous section the original signature of Th1, {Tbet}, can be extended by the refined signature {IFNγ, SocS1, Tbet}. In addition, this cell-type has a single variant, associated with refined signature {IFNγ, SocS1, Tbet, IL12R, STAT4}. This is in agreement with the biological role of the minor component STAT4 later confirmed in the literature. For instance, Thieu *et al* showed that STAT4 is required during the differentiation process of Th1 in order to achieve a complete phenotype [29]. On the left side of the lattice, the cell type Th2 (GATA3), associated with a single steady state has the refined signature {GATA3, IL4, IL4R, STAT6}.

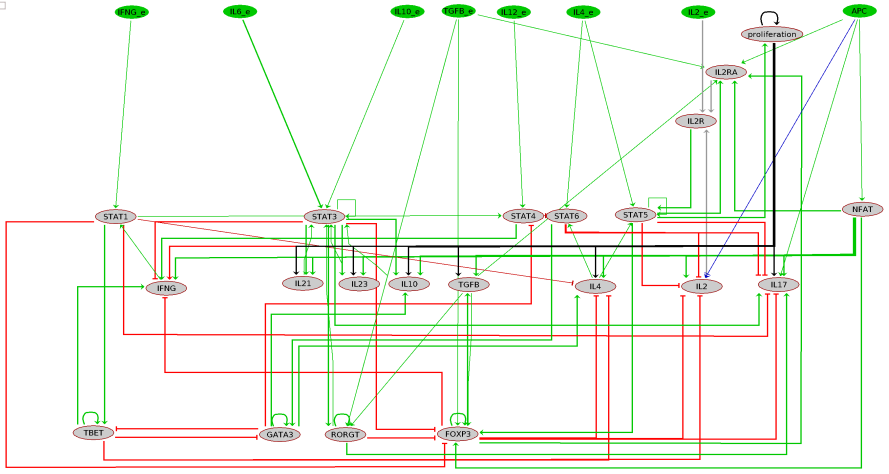
4.3. Comparing the impact of different simulation conditions

Let us notice, however, that this network is no longer suitable to study the other subtypes (Treg and Th17). Indeed, the two transcriptional factors that regulate Treg and Th17 (FOXP3 and RORγT) are not involved. Also, the network does not take into account the influence of other external stimuli implied in Treg and Th17 regulation. A larger network was defined in order to understand the role of the microenvironment [22]. More precisely, Naldi *et al* [22] have integrated transcriptional pathways to enrich the model of Th differentiation. The extended model encompasses 65 components and is controlled by 13 inputs representing external environmental stimuli (see Fig. 6(a)). In their study, the authors tested several input combinations (see Fig. 6(b)) and evidenced that the dynamical simulations generated 38 steady states (see Fig. 6(d) - Condition 1). For each phenotype (cell type) a signature was introduced corresponding to the expression of one master regulator of the network (Tbet, GATA3, FOXP3 or RORγT for, respectively, Th1, Th2, Treg and Th17) (Fig. 6(c)). Moreover, the authors set up the initial state which represents the Th0 type. One of the main results of Naldi et al.'s publication was a classification table for the steady states of the system which was derived from the pre-determined signatures. With the table, subtypes of Th cells were introduced and characterized by sub-patterns of expressed proteins.

Our methods and tool allowed us to perform several simulations of Naldi's network [22]. First, we used the input conditions introduced in [22] (8 environment conditions described in Fig. 6(b)) and generated the 37 steady states that can be reached from the Th0 initial state. The lattice generated by the FCA contained 59 concepts (Fig. 7). Among them, we checked that all hybrids identified in [22] effectively corresponded to a hybrid concept as defined in our formalism. As expected and shown in Fig. 6(e), the 38 steady states could all be automatically characterized to satisfy either a canonical signature (4 steady states for Th0, 12 steady states for Th1, 6 steady states for Th2, 1 steady state for Th17 and 1 steady state for Treg) or a hybrid signature (4 steady states for Th1-Th17, 1 steady state for Th1-Treg, 3 steady states for Th2-Th17, 2 steady states for Th17-Treg, 2 steady states for Th1-Th17-Treg and 1 steady state for Th2-Th17-Treg).

In order to study the influence of initial conditions on results, we simulated the network with the same input conditions (8 environmental stimuli) but by

a)



b)

No Stimulation	None (All = 0)
APC only	APC (APC = 1)
Pro-Th1 (1)	APC, $IL2_e$, $IFNg_e$
Pro-Th1 (2)	APC, $IL12_e$
Pro-Th2	APC, $IL4_e$, $IL6_e$
Pro-Th17	APC, $IL6_e$, $TGFb_e$
Pro-Treg (1)	APC, $IL2_e$, $TGFb_e$
Pro-Treg (2)	APC, $IL10_e$

c)

Phenotypes	Signatures
Th0	"empty"
Th1	Tbet
Th2	GATA3
Th17	RORGT
Treg	FOXP3

d)

	Condition 1	Condition 2	Condition 3
Input	Fig 5b	Fig 5b	All
Initial State(s)	Th0	All	All
Nb Steady States	37	55	435
Nb Canonical Classes	5	5	5
Nb Hybrid Classes	6	7	7
Nb Formal Concepts	59	116	789
Lattice graph available	Fig 6a	Fig 6b	No

Figure 6: Network controlling the differentiation of Lymphocyte T helper (Th) with the input environments used for the dynamics simulation and the signatures for the classification. Those data were described in [22]. (a) Reduced network of the differentiation of Th (35 components). Gray nodes correspond to internal compounds and green to the 13 inputs components. Activation regulations are represented with green arrows and indirect interactions resulting from the reduction are dotted arrows. Inhibitions are represented with red arrows. (b) Configuration of the input conditions used during simulation. Each row corresponds to one combination of inputs used for the simulation. (c) Initial signature for each cell type. Each row corresponds to one signature with the expression of one master regulators. The signature of Th0 is empty. However, some components in the system might be expressed. (d) Summary for the comparison of different simulation conditions of the network.

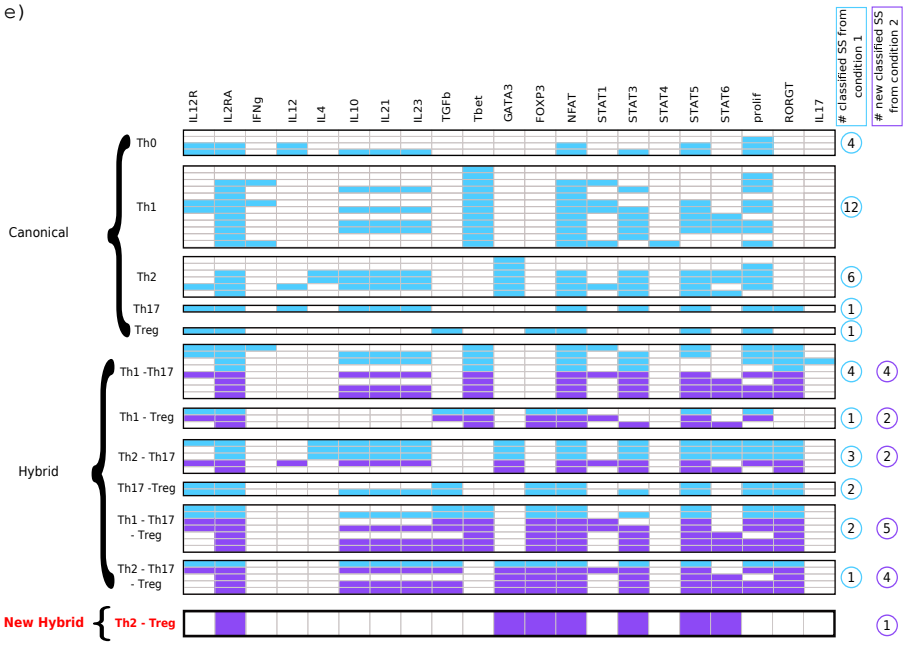


Figure 6: Network controlling the differentiation of Lymphocyte T helper (Th) with the input environments used for the dynamics simulation and the signatures for the classification. (e) Comparison between the classified steady states from condition 1 (blue) and condition 2 (purple) of the network controlling the differentiation of Lymphocyte T helper (Th)

relaxing the constrain on the Th0 initial state and allowing any state to be considered as an initial state. This novel simulation condition generated 55 steady states (Fig.6(d) - Condition 2), whose analysis produced the concept lattice shown in Fig. 8. This lattice contained 116 formal concepts, illustrating the strong dependence of this type of study on the conditions of network simulation and the increasing complexity of studying it because the signature compounds are drowned in the entire lattice. Interestingly, in comparison with the previous simulation, almost all 17 additional steady states are associated with an already known hybrid class (Purple lines in Fig. 6(e)): 4 novel steady states are associated with the hybrid cell type Th1-Th17, 2 with the hybrid cell-type Th1-Treg, 2 with the hybrid cell-type Th2-Th17, 5 with the hybrid Th1-Treg-Th17 and finallly 4 novel steady states with the hybrid cell-type Th2-Treg-Th17.

In contrast, one steady state could not be classified according to these canonical and hybrid cell-types. Based on our analysis, the hybrid type (Th2-Treg) is required to explain the data whereas there was no steady state associated with this phenotype in Naldi's work. Our prediction of this new hybrid has been in fact validated in the literature: Wang *et al* [33] worked on the role of $GATA3$ in the regulation of Treg function. The authors showed that the deletion of GATA3 expression induced an inflammatory disorder in mice with a decrease in FOXP3 expression. They also described that GATA3 can bind to a specific DNA sequence in the *FOXP3 locus* in the Treg cells. Our framework enables to understand better when this Th2-like Treg may occur. According to the simulations, this new hybrid phenotype appears only when the microenvironment is defined as pro-Th2, with activated APC, IL6 and IL4 input components.

4.3.1. A robust characterization of phenotypes

Comparing the two former simulation conditions highlights that extending the possibilities of initial states may have a large impact on the number of steady states and therefore on the lattice size, but it is only a low level vision of the cell behaviour. The number of hybrid for the other simulation condition does not change much : it increases only by 1 in our case. To push forward this idea and test the scalability of our method, we relaxed all the environmental conditions of [22] and simulated the network according to any initial state and any value for the environmental variables. This generated a family of 435 steady states (Fig. 6(d) - Condition 3). In this situation, the concept lattice contains 789 formal concepts. The graph is too large for being visualized with the *graphviz* package used in our method and we do not provide a figure for this lattice. However, our method evidenced that the numbers of canonical and hybrid classes remain the same as in the previous simulation, that is, 5 canonical classes and still 7 hybrid classes. This demonstrates that our method is robust to the study of large environments and it allows to certify that all possible behaviors in terms of variants were described in the previous simulations.

4.4. Classifying variants according to hybrids cell-types

Although the former analysis evidenced that the number of hybrid is relatively constant with respect to the simulation conditions of a Boolean network,

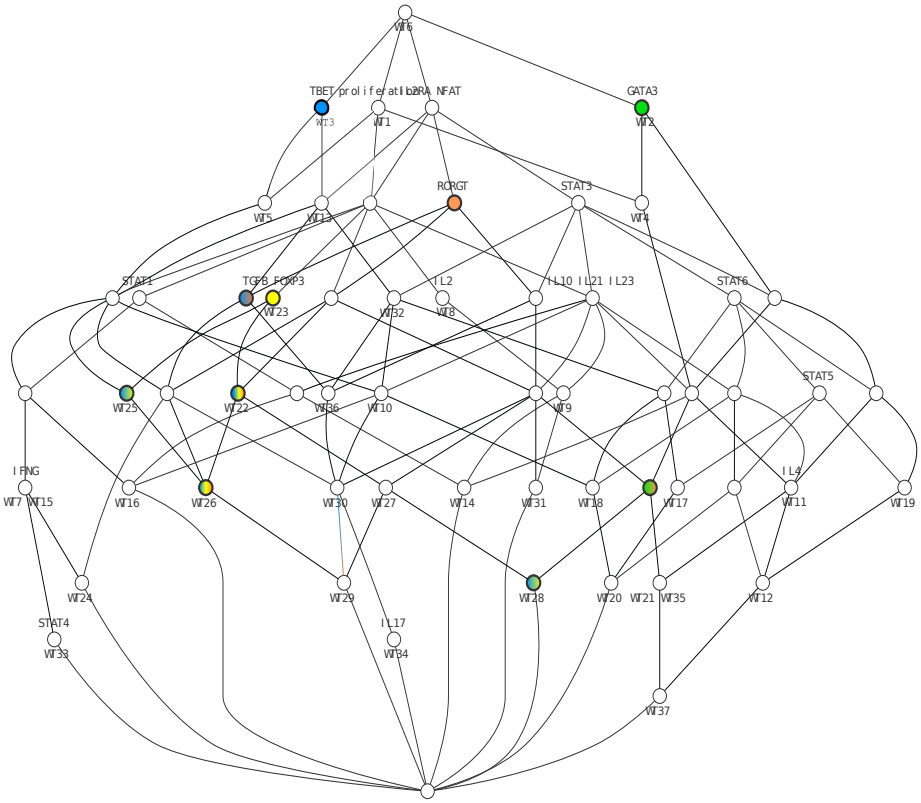


Figure 7: Lattice associated with a simulation of the network depicted in Fig. 6 with Th0 as initial state The lattice is built with the 37 steady states obtained by simulating 8 different input environments with Th0 as initial state. Each node corresponds to a concept (59 nodes). Each edge represents the inclusion relation between two sets of states or components of the concepts. The five concepts associated with canonical signatures (plain) and the 6 concepts associated with hybrid signatures (gradient) are depicted in bold.

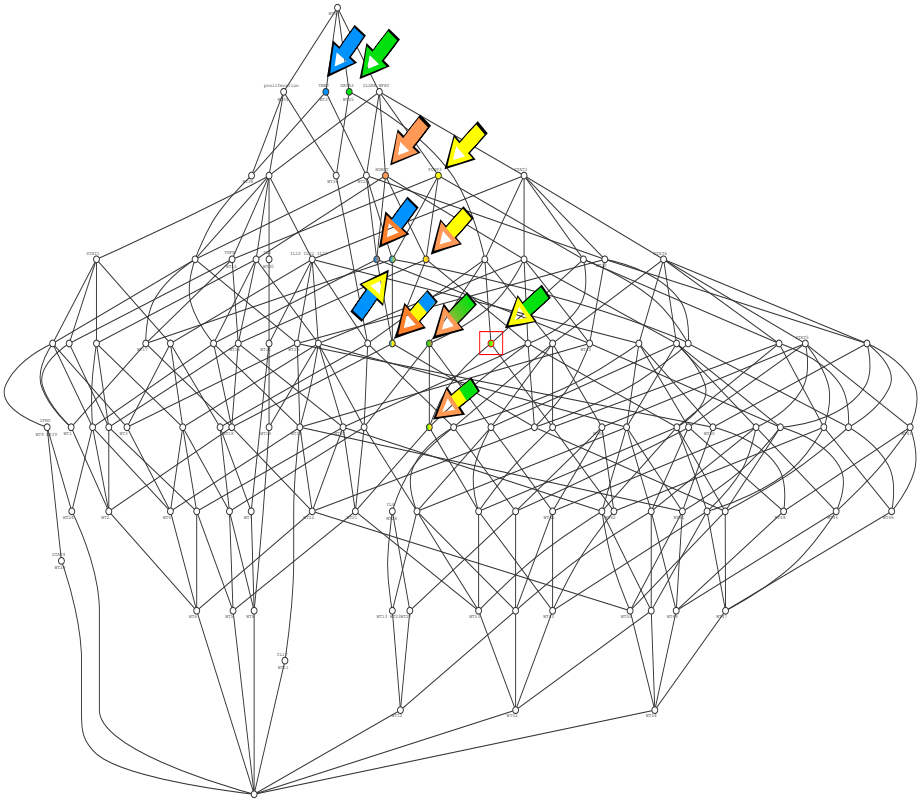


Figure 8: Lattice associated with a simulation of the network depicted in Fig. 6 with all possible values of internal genes or proteins as initial state. The lattice is built according to the 55 steady states obtained by simulating 8 different input environments with all possible values of internal genes or proteins as initial state. It contains 116 nodes. In comparison with the lattice shown in Fig. 7, we notice that the number of concepts has nearly doubled, concepts associated with canonical (plain) and hybrid phenotypes (gradient) are rather stable, since a unique additional hybrid concept is created in the lattice (red frame).

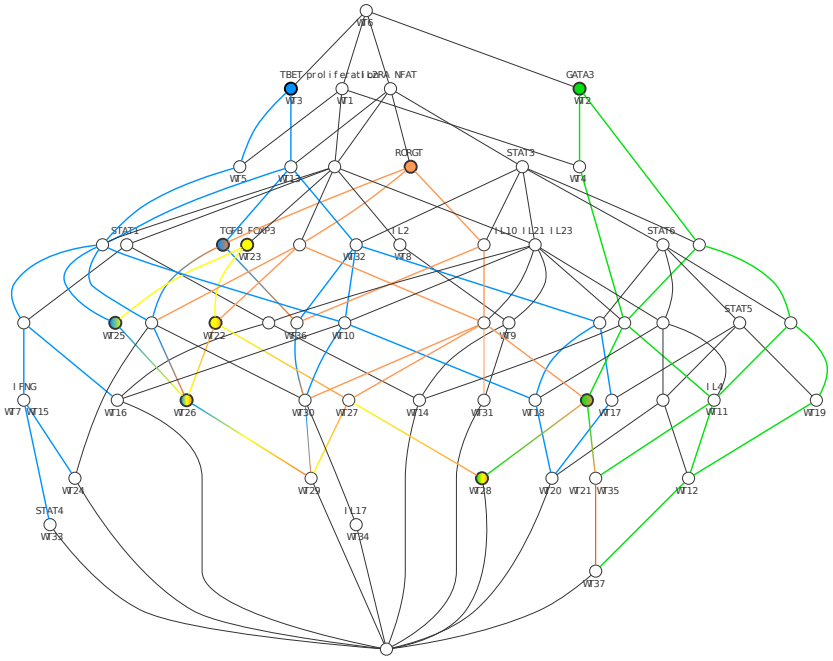
the number of variants depends on the number of formal concepts and therefore may be intractable when either the simulation conditions or the number of components of the network increase. To address this issue, we advocate the use of hybrids as a way to classify variants. Fig. 9 details the classification of steady states for the simulation of Naldi's network [22] according to 8 stimuli/environmental conditions and Th0 as initial condition. In Fig. 9(a), the different variants are classified according to the few canonical or hybrid concepts they are linked to. For instance, the formal concept associated with Th1 signature (TBET) has 37 variants. Among them, two variants are hybrid nodes associated with Th1-Th17, Th1-Treg and Th1-Th17-TReg. 11 of the Th1 variants are actually Th1-T17 variants. 3 of the variants are Th1-Treg variants. Among them, 3 variants constitute the hybrid Th1-Th17-Treg and its variants. This analysis highlights that canonical concepts and hybrids allow the lattice to be decomposed into classes of variants such that each variant which is neither a hybrid nor a canonical concept belongs to a single class. Fig. 9(b) provides a synthetic representation of the variant classes and the global structure between them.

5. Discussion

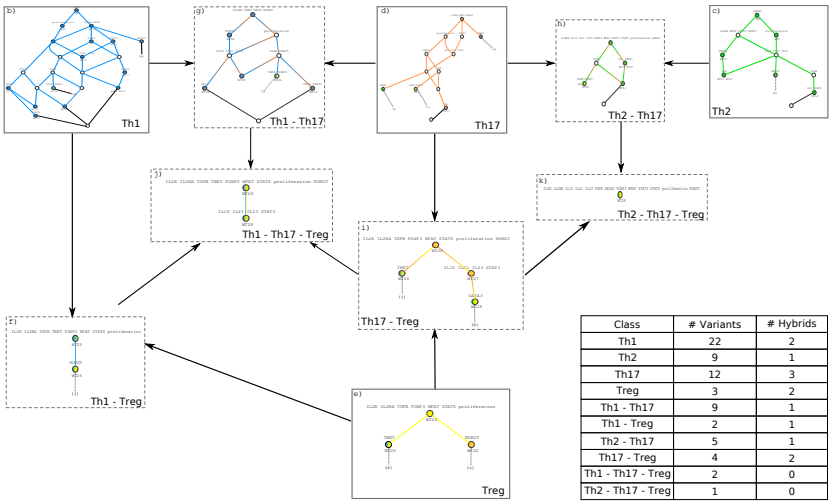
Data obtained from "omics" technologies provide a description of cellular compounds (gene, RNA, protein). Systems biology is based on this knowledge in order to analyze the dynamical behavior (phenotype) of the system in different conditions (specific environment or even mutations). When the number of simulation conditions increases, a major bottleneck of these analyses is the classification of the flood of steady states generated during the network simulation. Indeed, the main issue is that the system behavior is modeled by several phenotypes, each characterized by a few master regulators (genes, proteins or markers...). Distinguishing the activation of one or several master regulators in a large family of system's states is beyond the reach for standard clustering methods which are all subject to bias induced by their clustering metric.

To overcome this limitation, we promote the use of Formal Concept Analysis, a symbolic bi-clustering approach used in knowledge discovery and data mining. Our study suggests that FCA is accurate not only to extend expert-based signature according to the dependencies carried by the network dynamics, but also to automatically identify hybrid phenotypes associated with several signatures, together with initial conditions that may lead to new hybrid phenotypes. In addition, thanks to the hierarchy carried by the lattice structure of formal concepts, all the variants of canonical and hybrid phenotypes can be sorted in order to illustrate, for each phenotype, the role of biological compounds which are involved in signatures although they are not master genes. Such a distinction between master regulator and secondary regulator in signature was especially introduced in [2]; our method may provide a structure to systematically investigate the role of secondary master regulators in variant phenotypes.

To illustrate our approach, we studied several Boolean models for the gene regulatory system controlling the differentiation of LTh [26, 22]. The identifi-



(a)



Class	# Variants	# Hybrids
Th1	22	2
Th2	9	1
Th17	12	3
Treg	3	2
Th1 - Th17	9	1
Th1 - Treg	2	1
Th2 - Th17	5	1
Th17 - Treg	4	2
Th1 - Th17 - Treg	2	0
Th2 - Th17 - Treg	1	0

(b)

Figure 9: Analysis of the classification and identification of the hybrid classes based on the steady states generated from the condition 1 in Fig. 6(d) (a) Lattice with each canonical class in plain color (Th1 ● ,Th2 ● ,Th17 ● ,Treg ●) and hybrids classes. (b) Relation between each class based on the lattice. Each plain block is the sub-lattice of a canonical class. Each dashed block is the sub-lattice of a hybrid class. The nodes in blocks are the variants associated to the class.

cation of proteins involved in the transition from one phenotype to another are fully studied in cell plasticity or development of target therapy. In [22], the authors have specified one initial state for each simulation, the Th0 type because it is the naive, inactivated form of LTh. Several input sets have been used showing the microenvironment implication in the choice of sub-type differentiation. Our method resulted in the same association of steady states according to the different subtypes of LTh as in [26, 22]. The classification in [26, 22] was done manually which needed to classify one steady state at a time. But in our approach, all steady states are taken in one time. Moreover, by relaxing the constraints of the stimuli conditions and the simulation initiation, we have rationalized and systematized the study of phenotypes and evidenced that a new hybrid should be added to the family of considered phenotypes to complete the set of possible subtypes of the system.

The first limitation of our method is related to performance when the Boolean Network's size increases. In terms of complexity, the number of formal concepts increases exponentially with respect to the number of objects (here, model steady-states) or attributes (model nodes). The computation of the lattice structure (subconcept-superconcept relation) is also computationally demanding, but we use this information only at the very last step of our workflow when computing the subgraphs associated with each signature. Thanks to this strategy of strictly limiting the computation of the lattice structure, our approach scaled to the study of all steady states on the full network from [22] (65 nodes with 24,267 steady states). Anyway, when too large, the number of steady-states or nodes of the model is a limitation for this method. Some specialized tools such as *In-close*[1] are helpful to handle a larger amounts of concepts.

Another level of complexity is related to the reachability properties. Dynamical simulations are performed given an initial state and some inputs, in order to identify reachable steady states and attractors. For instance, in [22], the initial configuration is required to study the plasticity of Th cell types after the classification of all steady states according to phenotypes' signatures. It provides a very useful information about which combinations of initial states and input components lead to a specific phenotype. With our method, the Th2-Treg hybrid class shown in Fig.6(e), was not identified in the seminal paper because it is generated only when the initial input is pro-Th2 and the initial state is different from Th0, which was not tested in [22]. However, our method is only able to store this information if the search space is constrained enough, as in [22]. Addressing this issue will require for instance to develop technics based on model-checking for the identification of initial configurations leading to any steady states. When the initial state information is not needed, the sole identification of steady-states is less computationally demanding.

A second limitation of this method is to define the signature considering only the presence (activation) of biological components. Clearly, signatures of

[1] `https://sourceforge.net/projects/inclose/`

phenotypes may also consist in the inactivation of some biological components, always missing in the concepts associated to canonical signatures. For instance, in [2], the Th1 signature is defined by the inactivation of all the master regulators but the secreted cytokines (IL4 and IL17). The reason of limiting signatures definition to activated components is the fact that FCA generates the concepts according to the presence of attributes shared by objects. A strategy to overcome this issue is to expand the matrix used as input for FCA by duplicating each attribute (compound) v in v, *not* v with the implicit constraint that exactly one of them is present at a time: v + *not* $v = 1$. A complementary enrichment of the method could be to take into account continuous – or at least multiple – values, either derived from differential equations modelling, or corresponding to biological data with samples as object and the gene/protein as attribute. This extension could be done be relying on FCA tools that handle numerical tables.

Our method has been developed for the analysis of steady states. As described in Sec. 2.1, Boolean Networks also generate *cyclic attractors*, describing oscillations through several states of the system. An interesting extension is to consider these cyclic attractors, and associate to them one or several signatures. The identification of cyclic attractors is not an easy task. Moreover, they are composed of a -possibly very large- set of states, and it is not always obvious to express them in a compact way. To cope with these difficulties, some strategies are possible. Usually, we try to express the cyclic attractors by one or several "schemes". They are defined with abstract states with constant components (meaning that all the states gathered in the scheme have this component fixed to the same value, 0 or 1), and cycling components. If the stabilized components match a signature, we may associate this signature to the attractor.

Finally this method provides a way to measure to what extend a perturbation of the model (i.e. a mutation simulated by blocking a node to value 0 or 1) affects the system. Indeed, we can compare the signatures of the stable states obtained in the wild-type and a mutant model. Considering the mutation of a node of the network, its impact may be measured through a sort of robustness, depending on if the attractors obtained in the mutant model are assigned to the same phenotypes/signatures as the attractors of the wild-type (although different). Hence, a mutation may affect the dynamics in terms of loss or gain of phenotypes, loss or gain of reachability, etc...(e.g., loss of tumors suppressor in cancer cells. The impact of mutations on biological systems represented with logical networks was highlighted for instance in [25, 2]). Intuitively, we expect that the mutation of a master gene regulator in the signature of a phenotype will strongly impact the model. The definition of signatures in the context of gene deletion or ectopic still deserves to be further studied. A perspective is to formally model the effect of a perturbation as an operation over the concept lattice. The approach would require to compute the steady-states for the wild-type model and all the perturbed models to consider (for instance, models perturbed with at most two knock-out or ectopic perturbations). The modelling issue would then be to figure out how signature and hybrid can be defined on

formal concepts of this extended set of steady-states, by taking into account the type of perturbations that was performed to generate the considered steady states.

References

[1] Abou-Jaoudé, W., Monteiro, P. T., Naldi, A., Grandclaudon, M., Soumelis, V., Chaouiya, C., Thieffry, D., 2014. Model checking to assess T-helper cell plasticity. Frontiers in bioengineering and biotechnology 2, 86.

[2] Abou-Jaoudé, W., Ouattara, D. A., Kaufman, M., 2009. From structure to dynamics: Frequency tuning in the p53–Mdm2 network: I. Logical approach. Journal of Theoretical Biology 258 (4), 561–577.

[3] Alam, M., Le, T. N. N., Napoli, A., 2016. Latviz: A new practical tool for performing interactive exploration over concept lattices. In: Proceedings of the Thirteenth International Conference on Concept Lattices and Their Applications, Moscow, Russia, July 18-22, 2016. pp. 9–20.

[4] Barthélémy, J.-P., Guénoche, A., 1991. Trees and proximity representations. John Wiley & Sons.

[5] Bourneuf, L., Nicolas, J., 2017. Fca in a logical programming setting for visualization-oriented graph compression. In: Bertet, K., Borchmann, D., Cellier, P., Ferré, S. (Eds.), Formal Concept Analysis. Springer International Publishing, Cham, pp. 89–105.

[6] Caza, T., Landas, S., 2015. Functional and Phenotypic Plasticity of CD4(+) T Cell Subsets. BioMed research international 2015, 521957.

[7] De Jong, H., 2002. Modeling and simulation of genetic regulatory systems: A literature review. Journal of Computational Biology 9 (1), 67–103.

[8] de Sousa Abreu, R., Penalva, L. O., Marcotte, E. M., Vogel, C., 2009. Global signatures of protein and mRNA expression levels. Mol Biosyst 5 (12), 1512–1526.

[9] Fang, D., Zhu, J., 2017. Dynamic balance between master transcription factors determines the fates and functions of CD4 T cell and innate lymphoid cell subsets. The Journal of experimental medicine 214 (7), 1861–1876.

[10] Fauré, A., Naldi, A., Chaouiya, C., Thieffry, D., 2006. Dynamical analysis of a generic boolean model for the control of the mammalian cell cycle. Bioinformatics 22 (14), e124–e131.

[11] Folschette, M., Paulevé, L., Magnin, M., Roux, O., 2015. Sufficient conditions for reachability in automata networks with priorities. Theoretical Computer Science 608, Part 1, From Computer Science to Biology and Back, 66 – 83.

[12] Ganter, B., Stumme, G., Wille, R., 2005. Formal concept analysis : foundations and applications. Springer.

[13] Ganter, B., Wille, R., 1999. Formal concept analysis : mathematical foundations. Springer.

[14] Gonzalez, A. G., Naldi, A., Sánchez, L., Thieffry, D., Chaouiya, C., 2006. GINsim: A software suite for the qualitative modelling, simulation and analysis of regulatory networks. BioSystems 84 (2), 91–100.

[15] Ivanova, E. A., Orekhov, A. N., 2015. T Helper Lymphocyte Subsets and Plasticity in Autoimmunity and Cancer: An Overview. BioMed research international 2015, 327470.

[16] Kauffman, S., 1969. Metabolic stability and epigenesis in randomly constructed genetic nets. Journal of Theoretical Biology 22 (3), 437–467.

[17] Levy, N., Naldi, A., Hernandez, C., Stoll, G., Thieffry, D., Zinovyev, A., Calzone, L., Paulevé, L., 2018. Prediction of Mutations to Control Pathways Enabling Tumour Cell Invasion with the CoLoMoTo Interactive Notebook (Tutorial) . Frontiers in Physiology 9, 787.

[18] Liberzon, A., Birger, C., Thorvaldsdottir, H., Ghandi, M., Mesirov, J. P., Tamayo, P., 2015. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst 1 (6), 417–425.

[19] Lihonosova, A., Kaminskaya, A., 2014. Using formal concept analysis for finding the closest relatives among a group of organisms. Procedia Computer Science 31 (Complete), 860–868.

[20] Mitsi, E., Kamng'ona, R., Rylance, J., Solórzano, C., Jesus Reiné, J., Mwandumba, H. C., Ferreira, D. M., Jambo, K. C., 2018. Human alveolar macrophages predominately express combined classical M1 and M2 surface markers in steady state. Respiratory research 19 (1), 66.

[21] Mosmann, T. R., Coffman, R. L., 1989. TH1 and TH2 Cells: Different Patterns of Lymphokine Secretion Lead to Different Functional Properties. Annual Review of Immunology 7 (1), 145–173.

[22] Naldi, A., Carneiro, J., Chaouiya, C., Thieffry, D., 2010. Diversity and plasticity of Th cell types predicted from regulatory network modelling. PLoS Computational Biology 6 (9).

[23] Ouattara, D. A., Abou-Jaoudé, W., Kaufman, M., 2010. From structure to dynamics: Frequency tuning in the p53-Mdm2 network. II: Differential and stochastic approaches. Journal of Theoretical Biology 264 (4), 1177–1189.

[24] Polynikis, A., Hogan, S. J., di Bernardo, M., 2009. Comparing different ODE modelling approaches for gene regulatory networks. J. Theor. Biol. 261 (4), 511–530.

[25] Remy, E., Rebouissou, S., Chaouiya, C., Zinovyev, A., Radvanyi, F., Calzone, L., 2015. A Modeling Approach to Explain Mutually Exclusive and Co-Occurring Genetic Alterations in Bladder Tumorigenesis. Cancer Res 75 (19), 4042–52.

[26] Remy, E., Ruet, P., Mendoza, L., Thieffry, D., Chaouiya, C., 2006. From Logical Regulatory Graphs to Standard Petri Nets: Dynamical Roles and Functionality of Feedback Circuits. In: Priami, C., Ingólfsdóttir, A., Mishra, B., Riis Nielson, H. (Eds.), Transactions on Computational Systems Biology VII. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 56–72.

[27] Samaga, R., Von Kamp, A., Klamt, S., 2010. Computing combinatorial intervention strategies and failure modes in signaling networks. J. Comput. Biol. 17 (1), 39–53.

[28] Sheng, J., Chen, Q., Soncin, I., Ng, S. L., Karjalainen, K., Ruedl, C., 2017. A Discrete Subset of Monocyte-Derived Cells among Typical Conventional Type 2 Dendritic Cells Can Efficiently Cross-Present. Cell Reports 21 (5), 1203–1214.

[29] Thieu, V. T., Yu, Q., Chang, H.-C., Yeh, N., Nguyen, E. T., Sehra, S., Kaplan, M. H., 2008. Signal transducer and activator of transcription 4 is required for the transcription factor T-bet to promote T helper 1 cell-fate determination. Immunity 29 (5), 679–90.

[30] Thomas, R., 1991. Regulatory networks seen as asynchronous automata: A logical description. Journal of Theoretical Biology 153 (1), 1–23.

[31] Thomas, R., Thieffry, D., Kaufman, M., 1995. Dynamical behaviour of biological regulatory networks—I. Biological role of feedback loops and practical use of the concept of the loop-characteristic state. Bulletin of Mathematical Biology 57 (2), 247–276.

[32] Vignali, D. A. A., Collison, L. W., Workman, C. J., 2008. How regulatory T cells work. Nature reviews. Immunology 8 (7), 523–32.

[33] Wang, Y., Su, M. A., Wan, Y. Y., Sep 2011. An essential role of the transcription factor GATA-3 for the function of regulatory T cells. Immunity 35 (3), 337–348.

[34] Wucher, V., Tagu, D., Nicolas, J., 2015. Edge selection in a noisy graph by concept analysis: Application to a genomic network. In: Lausen, B., Krolak-Schwerdt, S., Böhmer, M. (Eds.), Data Science, Learning by Latent Structures, and Knowledge Discovery. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 353–364.

[35] Yagil, G., 1975. Quantitative aspects of protein induction. In: Horecker, B., Stadtman, E. (Eds.), Current topics in Cell regulation. Academic Press, pp. 183–237.

[36] Yepes, S., Torres, M. M., Andrade, R. E., 2015. Clustering of Expression Data in Chronic Lymphocytic Leukemia Reveals New Molecular Subdivisions. PloS one 10 (9), e0137132.