Modelling cell lineage using a meta-Boolean tree model with a relation to Gene Regulatory Networks

Jan-Å Ke Larsson, Niclas Wadströmer, Ola Hermanson, Urban Lendahl, Robert Forchheimer

www.elsevier.com/locate/yjtbi

Modelling cell lineage using a meta-Boolean tree model with a relation to Gene Regulatory Networks

Jan-Å ke Larsson, Niclas Wadströmer, Ola Hermanson, Urban Lendahl, Robert Forchheimer

PII: S0022-5193(10)00531-X
DOI: doi:10.1016/j.jtbi.2010.10.003
Reference: YJTBI 6187

To appear in: *Journal of Theoretical Biology*

Received date: 3 December 2009
Revised date: 31 August 2010
Accepted date: 4 October 2010

Cite this article as: Jan-Å ke Larsson, Niclas Wadströmer, Ola Hermanson, Urban Lendahl and Robert Forchheimer, Modelling cell lineage using a meta-Boolean tree model with a relation to Gene Regulatory Networks, *Journal of Theoretical Biology*, doi:10.1016/j.jtbi.2010.10.003

Modelling cell lineage using a meta-Boolean tree model with a relation to Gene Regulatory Networks

Jan-Åke Larsson[a], Niclas Wadströmer[a], Ola Hermanson[b], Urban Lendahl[b], Robert Forchheimer[a]

[a] *Department of Electrical Engineering, Linköping University, Sweden*
[b] *Karolinska Institutet, Stockholm, Sweden*

Abstract

A cell lineage is the ancestral relationship between a group of cells that originate from a single founder cell. For example, in the embryo of the nematode *Caenorhabditis elegans* an invariant cell lineage has been traced, and with this information at hand it is possible to theoretically model the emergence of different cell types in the lineage, starting from the single fertilized egg. In this report we outline a modelling technique for cell lineage trees, which can be used for the *C. elegans* embryonic cell lineage but also extended to other lineages. The model takes into account both cell-intrinsic (transcription factor-based) and -extrinsic (extracellular) factors as well as synergies within and between these two types of factors. The model can faithfully recapitulate the entire *C. elegans* cell lineage, but is also general, i.e. it can be applied to describe any cell lineage. We show that synergy between factors, as well as the use of extrinsic factors, drastically reduce the number of regulatory factors needed for recapitulating the lineage. The model gives indications regarding co-variation of factors, number of involved genes and where in the cell lineage tree that asymmetry might be controlled by external influence. Furthermore, the model is able to emulate other (Boolean, discrete and differential-equation-based) models. As an example, we show that the model can be translated to the language of a previous linear sigmoid-limited concentration-based model (Geard and Wiles, 2005). This means that this latter model also can exhibit synergy effects, and also that the cumbersome iterative technique for parameter estimation previously used is no longer needed. In conclusion, the proposed model is general and simple to use, can be mapped onto other models to extend and simplify their use, and can also be used to indicate where synergy and external influence would reduce the complexity of the regulatory process.

Keywords: Differentiation, Transcription factor, Asymmetric cell division

1. Introduction

1.1. Cell lineage and control of cell differentiation

All cells in the adult individual are derived from the fertilized egg cell, and are thus ancestrally related in the organism. In the cell lineage, cells undergo differentiation to various cell types and to understand the principles for this progression from uncommitted to specialized cells is of importance. The differentiation of cells from pluripotent to specialized cell types is controlled by the genes in the genome of each cell. With very few exceptions, the genome, i.e., the collection of genes in the chromosomes, is the same in all cells, but what endows a cell with its unique characteristics is the subset of the total number of genes that is activated in each cell type (Lécuyer and Tomancak, 2008).

Both cell-intrinsic and -extrinsic mechanisms control which genes to express in a given situation (Bannister and Kouzarides, 2005). On the cell-intrinsic side, an important regulation of genes is executed by transcription factors, DNA-binding proteins that control gene expression by binding to regulatory regions (enhancers and/or promoters) for the different genes, and which can turn the gene on or off. The degree of compaction and accessibility of the DNA in the chromosomes is also a critical component in gene regulation, and the DNA can be directly modified by methylation but also subjected to regulation by histone proteins, which bind to DNA in chromatin and in turn can be modified by epigenetic mechanisms. Additional control levels for gene expression are regulation of the stability of mRNAs or proteins. A key aspect of cell differentiation is also regulation by asymmetric cell division, where one cell divides to generate two distinct daughter cells by localizing specific proteins, asymmetric determinants, to only one of the daughter cells.

Cell-extrinsic modes of gene regulation involve communication between cells at various distances and levels. Cell-cell communication mechanisms range from direct cell-cell interaction mechanisms to more long-range influences, exerted for example by secretion of ligands from one cell that bind to and activate cell-bound receptors on other cells, followed by transmission of these signals into gene regulatory events in the cell nucleus. When invoking cell-extrinsic mechanisms, it is also important to consider the spatial organization of cells in the organism: cells located next to each other can engage in direct cell-cell communi-

cation, whereas more distantly related cells can contribute by production of diffusible factors.

When modelling cell lineage it is important to have detailed information about a defined cell lineage with regard both to cell division and the resulting cell types. Given the size and complexity of most multicellular organisms, tracing of entire cell lineages has proven difficult and has only been achieved in a few, select cases. The most well understood cell lineage is derived from the nematode *Caenorhabditis elegans* (*C. elegans*). This worm is usually a hermaphrodite (there are males, but they are rare), in the adult stage approximately one millimetre long, and composed of less than 1000 cells, not counting the germ cells. Due to its transparency, it has been possible to use light microscopy to monitor every cell division in the developing nematode, and thus deduce the complete embryonic cell lineage (Sulston et al., 1983; see also `www.wormatlas.org` for images of the worm and a depiction of the entire cell lineage tree). More recently, DNA sequencing of the *C. elegans* genome has revealed that it contains approximately 20000 genes.

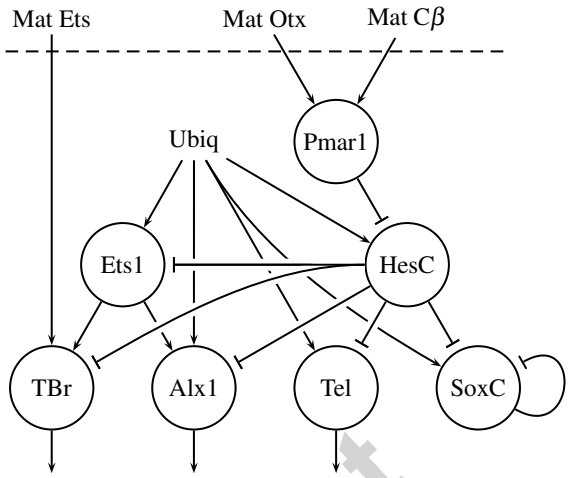

Figure 1: Part of gene regulatory network for the specification of the skeletogenic micromere lineage of the sea urchin embryo (modified from Oliveri et al., 2008).

1.2. Models of dynamic gene regulatory networks

A traditional way to model the interactions between genes in the cell is the so called Gene Regulatory Network (GRN) description. This is a (directed) graph description in which the vertices correspond to genes and edges to transcription factors. Edges are also assigned with a symbol to express whether transcription factors act as activators (+ or a normal arrowhead ↓) or repressors (− or a bar as arrowhead ⊥), see Fig. 1. The *synergy* between these transcription factors (whether they combine in a linear or nonlinear way) is usually defined separately. Examples of recent GRN models are Platzer and Meinzer (2004) and Oliveri et al. (2008). Although such a graph gives a very informative map of gene relations it does not tell about the dynamics of the system, and therefore has limitations in its use.

Early models that included dynamics started to appear in the 1950s and 1960s. These models were differential equation models (inspired by the chemical model of Turing, 1952) and binary (Boolean) models (Kauffman, 1969; Thomas, 1973). Later, also other types of models such as rule-based and stochastic models were proposed, see de Jong (2002) and Smolen et al. (2000) for excellent reviews. Recent examples of modelling work include Platzer and Meinzer (2004), Geard and Wiles (2005), and Lohaus et al. (2007). A Boolean model is characterized by genes that are either on or off, and the expressed transcription factors are present or not at any specific instance in time. The set of genes (or transcription factors) are termed *state variables* and are collected into a *state vector*. In the differential equation models state variables take on continuous levels. Although other types of models have been developed, the Boolean models and differential equation models are still the main types studied today.

The models are generally nonlinear and will therefore require numerical evaluation. Being dynamic they describe the progression of the state variables over time. These variables typically represent the concentration of factors (for example, proteins). In the most simplistic case, the variables are binary valued and the "activation" of a gene and the corresponding expression of the protein is considered to be one and the same event (Kaletta et al., 1997; Platzer and Meinzer, 2004). In reality, levels of mRNA and protein expression do not always co-segregate, for example as a result of post-translational regulation of the mRNA. Protein and mRNA concentrations can thus vary over a wide range of values and most models today therefore use either multilevel or continuous values to represent them (Thomas et al., 1995; Bodnar, 1997; Bernot et al., 2004). When binary models are extended in this way they are referred to as multilevel Boolean models. Synergy relations, which in the binary case can be described by Boolean expressions (see e.g. Materna and Davidson, 2007; Knabe et al., 2008) become more complex though as there will be many more possible factor combinations (see Schilstra and Bolouri, 2003, for a discussion on various cis-regulatory functions).

An important issue concerns the representation of the time axis. Time may either be assumed to be discrete or continuous. Discretization may be for a purely mathematical or computational reason (time is sampled at sufficiently high rate to capture any nuance of the measured data), or the sampling interval is matched to some natural period of the modelled system, e.g., cell divisions.

Typically, binary and multilevel models use discretized time while differential equation models use continuous time. However, in the latter case it is straightforward to translate the differential equation into a discrete time difference equation. It is less common to combine binary (or multilevel) state variables with continuous time although some of the earlier proposals were actually such "automata theoretic" models (Thomas, 1973, 1978) as well as more recent

proposals such as that of Siebert and Bockmayr (2008).

Several authors have attributed the seemingly robust behaviour of the error-prone biological process to the fix-point property found in some non-linear feedback systems (Plahte et al., 1994; Kaneko, 1997; Furusawa and Kaneko, 1998b,a; Silva and Martins, 2003; Yoshida et al., 2005; Mochizuki, 2008). By this property is meant that the system will eventually reach a final stable state or a cycle of states (an attractor) even if the initial conditions are changing or there has been some perturbation during the process. Depending on the state definition such a fix-point could correspond to a fully specialized cell or even the formation of the whole organism. The proposal is elegant but it is usually a non-trivial task to find the non-linear system which has a set of predefined fix-points.

A crucial issue in cell lineage modelling concerns the mechanism behind asymmetric division. A simple assumption is that there is a specific factor ("asymmetric determinant") produced prior to cell division and found only in one of the progeny cells, (see e.g. Geard and Wiles, 2005). Such a simple mechanism is however too restricted to allow modelling of general lineage trees. Instead, two different factors can be used, one for each progeny cell. A yet more advanced model may use different asymmetric factors at different locations in the tree. This is strictly not necessary but could lead to more plausible descriptions from a biological point of view (Jan and Jan, 1998).

Some works address the relations between cell lineage and (3D) morphology. If morphological information is available it can be incorporated into the model, e.g., to control the influence of extrinsic factors or morphology (Bodnar, 1997; Bodnar and Bradley, 2001; Platzer and Meinzer, 2004; Smith et al., 2007). Alternatively, morphological features such as shape or colouring may be predicted from the model (see e.g. Furusawa and Kaneko, 2000; Silva and Martins, 2003).

Although most studies today are focused on differential equation approaches, there seems to be a renewed interest in Boolean models. The rationale is that quantitative knowledge of concentrations is usually not possible to obtain. Furthermore, the computational complexity of a Boolean model is lower, making it possible to evaluate reasonably large systems of cells. Examples of recent publications are Silva and Martins (2003); Bernot et al. (2004); Platzer and Meinzer (2004) and Siebert and Bockmayr (2008). Dedicated software packages have been developed to support the various types of models (Braun et al., 2003; Albert et al., 2008). Analysis techniques from the electronic field is also applicable as discussed in Dubrova et al. (2005). Evaluating a model to see how well it describes a biological function can be made in many ways. Some models are "high-level" in the sense that they only attempt to explain basic phenomena such as oscillatory behaviour or statistical distribution of factors or cell types (Furusawa and Kaneko, 1998a, 2000; Banzhaf, 2003), other are closer to known biological processes (Platzer and Meinzer, 2004) or even specific organisms (Bodnar and Bradley, 2001; Davidson, 2006). Related work that aims at establishing links between gene products and factor contents in the various cells is also essential to tune the models to approach "real life" (Bodnar and Bradley, 2001; Howard-Ashby et al., 2006; Wei et al., 2006).

1.3. Modelling the Caenorhabditis elegans cell lineage

The embryonic *C. elegans* cell lineage serves as a good starting point to model cell lineages, and there are several modelling strategies that can be considered for this type of study. Notably, Geard and Wiles (2005) applied a state-continuous, time discrete model. They used additive (linear) functions to describe both the promotor relationships as well as the current levels of activation of the genes. The latter are passed through a sigmoid function to constrain between 0 (not active) and 1 (fully active). They treat asymmetric cell division by a "relative position" input that is not considered part of the internal state, which is asymmetrically distributed, always to the right-hand daughter. This input, along with the gene activation in the mother cell, is used to calculate the gene activation in the daughter cells. A similar model can be found in Lohaus et al. (2007), with values ranging from -1 to $+1$ and where the asymmetric input only influences one specific gene.

A different modelling approach is used in Lee et al. (2008). Here, a probabilistic (Bayesian) network is used to build relations between gene perturbations and tissue-type changes. Based on known genotype/phenotype relations the network was able to accurately predict phenotypical result of non-trained mutations.

In another report, a Boolean model is introduced to describe the first two steps in epigenesis, namely determination of polarity and blastomere fate (Platzer and Meinzer, 2004). The determination of organ identity and morphogenesis is not part of the model. Although the model only uses binary-valued concentrations, the authors manage to show close agreement between modelled behaviour and the cell lineage deduced in vivo. While the aim is to closely follow the available transcription factors (which are based on real protein concentrations measured in the cells at varying states in the developing phase) at some places in their model, they have to introduce "pseudo-genes" to account for inconsistencies in their model. Many of these inconsistencies seem to be due to the lack of multiple-valued concentrations as well as lack of a mechanism for automatic decay of protein concentrations.

Finally, Azevedo et al. (2005) use a Boolean rule-based model to measure cell lineage complexity. They compare the complexity of the *C. elegans* lineage and three other metazoan lineages, and also use the measure on random lineages. They find that real lineages are simpler than the corresponding random lineages and discuss a number of reasons for this. They conclude with the suggestion that cell positioning and number poses a constraint that leads to this behaviour.

The aim of the present study is to produce a model that can generate the *C. elegans* embryonic cell lineage,

but that will also have the power to generate cell lineage trees in general. We aim to include a number of important characteristics from the cell lineage in our model. First, cell divisions occur at definite moments in time, which could be used as a natural discretization; this is naturally represented in the tree itself. Second, we obviously need means to account for factor content in the cells and its production (e.g., transcription). Third, we need a mechanism for symmetric and asymmetric cell division. Fourth, the various types of differentiated cells (neurons, muscle cells etc) should be represented in the model. Fifth, not only the repertoire of transcription factors, but also their concentrations can influence the gene regulatory output. Sixth and last, we explore how the use of cell-extrinsic influence as well as cell-intrinsic regulation including synergy effects that affect the complexity of the model.

All these aspects have been taken into consideration when establishing the cell lineage model. We aim to keep model complexity (relatively) low, without losing descriptive power. This is handled through a modified Boolean model, here denoted *meta-Boolean model*, which is able to handle synergy, decay of concentrations, as well as cell-cell signaling. Our hypothesis is that, while a traditional Boolean model is general, our tweaked (meta-Boolean) model is biologically plausible, and has several features that makes it useful in the context of currently used models. In particular, it can be used to investigate whether the inclusion of cell-cell signaling reduces the complexity of the resulting lineage descriptions, which would enable a more detailed analysis of, e.g., the suggestion of Azevedo et al. (2005).

2. Model construction

We have chosen to model the presence of factors, or their concentration, not the activity of genes as is the usual interpretation of a GRN. It is therefore important to note that we use the notion of Factor intentionally to refrain from talking about proteins or other specific biological substances as such. The factors used here are intended to constitute a general framework describing proteins as well as other substances (e.g., calcium ions), external influences like cell-to-cell signalling, or even purely physical factors like external pressure. This is similar to, but slightly more general in scope than the use of BICs (Biological Information Carriers) in Platzer and Meinzer (2004). Further, we intentionally use the notion of Regulator instead of Gene, again aiming for a general framework where regulators describe a process where presence of one factor leads to later presence of another, rather than the actual physical processes involved, say in transcription, synthesis, or degradation.

2.1. Construction of a meta-Boolean model

We must stress that we do not intend to model the dynamics of, for example, transcription factors at very low concentration or their binding to the promotor in the genetic transcription regulation. Nor are we intending to find steady states of the genetic regulation system. Instead, we intend to systematize dependencies and control of cell differentiation and division at a more coarse-grained level, governing the emergence of a cell lineage tree. We therefore start by using a Boolean representation with yes/no assignments to statements of the form "factor X has concentration in the interval 1 to 2 units." In what follows, we will first use a purely Boolean model, but will later extend this to a model with multiple levels.

It is sometimes argued that a Boolean model is less powerful than a model with continuous levels (e.g., Smolen et al., 2000), but as we shall see, for the present purposes, this is not the case. In fact, when time is discretized, a model with quantized levels is mathematically equivalent to a model with continuous concentration levels. This is because it is possible to enumerate the possible levels at each time in the continuous-level model and use these as quantized levels in the fully discrete model. Also, we concentrate on a deterministic model here, and note that probabilistic effects can be included at a later step using, for example, Monte-Carlo methods.

The main aim of the present paper is to model cell lineage, so the process of cell division must be included in the model. This is usually not done in the gene-regulatory-network (GRN) models that are common in the field, but is not difficult to include. For example, in Geard and Wiles (2005), cell division is the basis for the time discretization of the model: at every time-step, there is a cell division. Here, a slightly more general approach is used, by assigning a special factor or factors that initiate cell division. This can be interpreted as a corresponding biological process where a special factor keeps the cell cycle running, which is in line with the current understanding of how the cell cycle actually is controlled. Another process that needs to be included is asymmetric cell division. There are several biological processes that results in asymmetric cell division. For example, in the Drosophila embryo a particularly well characterized asymmetric process invokes the asymmetric determinant Numb, which is localized to one of the two daughter cells in an asymmetric cell division (Gönczy, 2008). It is quite simple to include this in the model, and the notation is described below.

The choice of notation may seem to be a minor issue, but an appropriate choice of notation is very helpful in solving a given problem, while an inappropriate choice may hide simple solutions from view. Therefore we have chosen to deviate from the usual notation, by not giving the map from one time-step to the next as a matrix, as is usually the case in mathematical representations of dynamic GRNs. Our choice here will instead be a rule set on the form "factor A will be present in the next time-step if factor B is present in the current" (a discussion of rule-set formalisms can be found in de Jong, 2002).

In most dynamic GRN frameworks, the map from one set of concentrations to the updated set at a later time

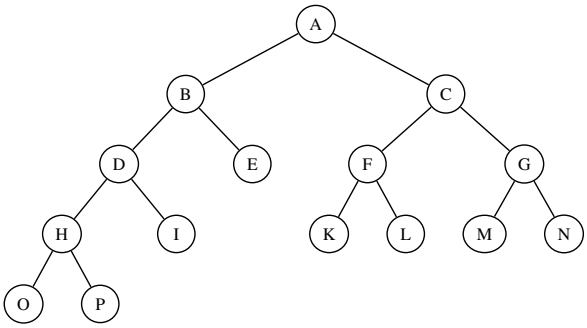

Figure 2: An example tree in canonical form. The root node contains the factor A, and the leaves contain E, I, K, L, M, N, O, and P, respectively.

is taken to be linear, limited by a sigmoid function (see de Jong, 2002, for a discussion of sigmoid functions in GRN models). This seems to be a severe restriction because it seems to prohibit synergy effects. In the present model, we explicitly allow synergistic (multiplicative) influence of factors, which will enable simplifications in the model in the form of reductions in the needed number of distinct factors, and in the needed number of regulator expressions. We will also see, in Section 5.2, that this is already present in some GRN frameworks, even though it is not explicitly visible in the formalism.

2.2. Factor content, transcription, and synergy

We start by defining a mathematical nomenclature to formally describe a cell lineage tree, factor content, and factor production (for example, transcription). A tree is a well-studied object in mathematics, and is in our case represented by the cell lineage tree. The cells are called nodes and the connections between a node and its daughters are known as edges. The terminal cells are denoted leaves, and the top cell in the tree (the zygote) is called the root of the tree. The order of the daughters is normally not important in the mathematical literature, but in a biological system the order may be of importance. In the case where the order is not important, there are several ways to draw a tree, and there is in this case a standard way to sort the branches so that the deepest branches are to the left to the tree. This is known as the canonical form of the tree (see Fig. 2).

Our initial choice of discretizing time at cell division is clearly visible in the tree. Each cell has a "factor content" that describes the properties of the present cell. Of course, nothing prevents a more fine-grained discretization of time, in which case a time-step might not correspond to a cell division, as indicated in Fig. 3a. Our initial choice of Boolean factors is also visible: in the root, the factor named A is present and it is not present in any of the following daughters. The regulators of our model produce factors that are present in the next time-step, and factors that are not produced disappear.

Figure 3: a) The regulator $g(\mathrm{B}|\mathrm{A})$ denotes that B is produced if A is present. b) The regulator $g(\mathrm{D}|\mathrm{B},\mathrm{C})$ denotes that D is only produced if *both* B and C are present.

To describe the production of factors we have chosen to use regulator expressions to establish rules on the form "Factor B will be produced if factor A is present" and the notation we use is

$$g(\mathrm{B}|\mathrm{A}).$$

The vertical bar is borrowed from conditional expressions in probability theory where it is read explicitly as "given that." The expression would then read "Given that factor A is present, factor B will be produced".

This notation makes it simple to write down expressions for synergy effects, where two factors are needed to produce a third:

$$g(\mathrm{D}|\mathrm{B},\mathrm{C}),$$

where the factor D would be produced only if both B and C are present, see Fig. 3b.

2.3. Symmetric and asymmetric cell division

We now need to model cell division by assigning a mechanism in the model that initiates cell division. One could use one specific factor that initiates cell division, but since we want to be able to model asymmetric cell division, we have chosen to have two specific factors a and b, with the following characteristics: *i*) the factors are always produced simultaneously by rules on the form

$$g(\mathrm{a},\mathrm{b}|\mathrm{E}),$$

and *ii*) the factors a and b immediately induce cell division upon production and disappear immediately after cell division. The process as represented within the model is indicated in Fig. 4a. This has bearings to proteins in the real, biological cell cycle, where levels of certain proteins such as cyclins oscillate during the different phases of the cell cycle.

It is important to stress that the individual factors a and b in our model are not intended to correspond to actual proteins or other chemical substances. Instead they serve as formal markers of the physical asymmetry appearing in asymmetric cell division. Such a system lends support from biology, as in the above described example where asymmetric distribution of the protein Numb plays a key role in the asymmetric cell division process in the Drosophila embryo (Gönczy, 2008). In this case, the factor Numb is uniformly produced in the mother cell, is asymmetrically distributed during the cell division process, and can subsequently be found in only one of the two daughter cells.

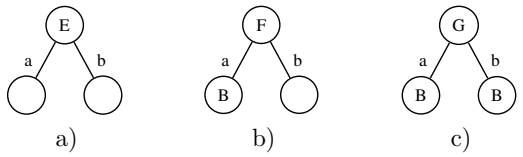

Figure 4: a) Cell division as represented in the model by the regulator $g(\mathrm{a},\mathrm{b}|\mathrm{E})$. b) Asymmetric cell division using the regulators $g(\mathrm{a},\mathrm{b}|\mathrm{F})$ and $g(\mathrm{B}|\mathrm{F},\mathrm{a})$. c) Symmetric cell division using the regulators $g(\mathrm{a},\mathrm{b}|\mathrm{G})$, $g(\mathrm{B}|\mathrm{G},\mathrm{a})$ and $g(\mathrm{B}|\mathrm{G},\mathrm{b})$; alternatively using $g(\mathrm{a},\mathrm{b}|\mathrm{G})$ and $g(\mathrm{B}|\mathrm{G})$.

Thus, the mechanism required for asymmetric cell division would be as follows: production of a factor which is present only in one of the daughter cells is described by an expression on the form

$$g(\mathrm{B}|\mathrm{F},\mathrm{a}),$$

which represents a process where production of the factor B is controlled by presence of the factor F, but after cell division the factor B is only present in one of the daughter cells, and not in the other (unless it is produced through another rule), see Fig. 4b.

In the case of symmetric cell division, both daughters receive the same content, and the factors a and b are only needed to account for the division itself as indicated in Fig. 4c. In what follows, the indices a and b on the edges will be suppressed, and the distinction will be left-right in the drawn tree instead.

Remarkably, with the regulator functions that are described in Fig. 4b, all cell lineage trees can be generated. Mathematically, this is because every division is given by regulators like that in Fig. 4a, and the content of every daughter as in Fig. 4b. This is true even in the absence of synergy; the asymmetric division mechanism of Fig. 4b is all that is needed. There is thus a one-to-one correspondence between a given tree and a list of regulator expressions (and a start node), see the Appendix for a formal proof.

To construct a given tree, we give each node a name and identify the name with a (formal) factor. Now if the node "A" is not a leaf (if it divides) like in Fig. 2, add a division regulator expression $g(\mathrm{a},\mathrm{b}|\mathrm{A})$. For each daughter, add a regulator expression that produces the content of it, in Fig. 2 the result is $g(\mathrm{B}|\mathrm{A},\mathrm{a})$ and $g(\mathrm{C}|\mathrm{A},\mathrm{b})$. For each of the produced nodes, continue this process; in Fig. 2 this results in

$$\begin{aligned}&g(\mathrm{a},\mathrm{b}|\mathrm{A}), g(\mathrm{B}|\mathrm{A},\mathrm{a}), g(\mathrm{C}|\mathrm{A},\mathrm{b}), g(\mathrm{a},\mathrm{b}|\mathrm{B}),\\ &g(\mathrm{D}|\mathrm{B},\mathrm{a}), g(\mathrm{E}|\mathrm{B},\mathrm{b}), g(\mathrm{a},\mathrm{b}|\mathrm{C}), g(\mathrm{F}|\mathrm{C},\mathrm{a}),\\ &g(\mathrm{G}|\mathrm{C},\mathrm{b}), g(\mathrm{a},\mathrm{b}|\mathrm{D}), g(\mathrm{H}|\mathrm{D},\mathrm{a}), g(\mathrm{I}|\mathrm{D},\mathrm{b}),\ldots\end{aligned}$$

For the example in Fig. 2, there will be 7 divisions and 14 content specifications, in total 21 regulator expressions, and to generate the whole tree, the start node and its content is also needed.

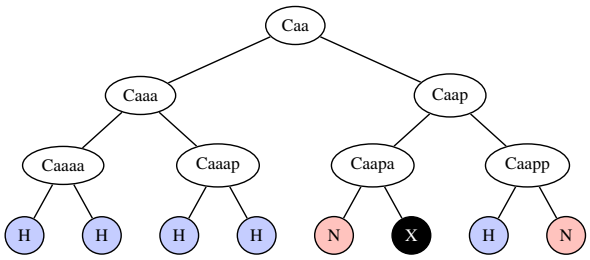

Figure 5: Part of the *C. elegans* lineage tree. The terminal cell types are hypodermis (H), nerve (N) and programmed cell death (X). The intermediate nodes are labeled using the established nomenclature for *C. elegans* ("a" for "anterior" and "p" for "posterior" denotes the post-division cell position relative to the orientation of the embryo, see e.g., Sulston et al., 1983). The cell names are here used simply as names of the node-specific factors.

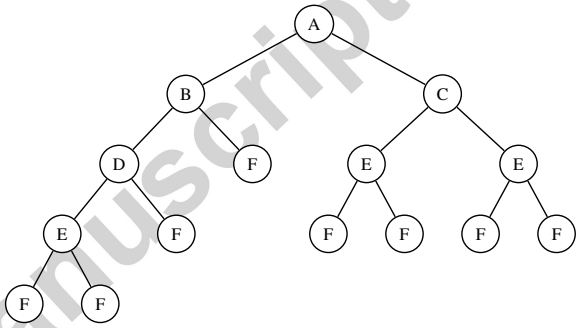

Figure 6: The example tree from Fig. 2 using all the possible reductions from identifying identical subtrees.

2.4. Cell differentiation and repression

In a biological cell lineage tree, many terminal cells are of the same type, and are really not unique in the manner suggested in Fig. 2. If we allow several cells in the tree to share "name," or really factor content, we can directly model cell differentiation. As an example, we can directly model a part of the *C. elegans* lineage tree, see Fig. 5 (more on modelling *C. elegans* in Section 3 below). The factor assignment is simple and there are 10 factors, 7 division regulator expressions, and 14 factor-specific regulator expressions, for example $g(\mathrm{H}|\mathrm{Caaaa},\mathrm{a})$.

Now, returning to the example in Fig. 2 but having identical terminal cells (leaves), i.e., only considering the form of the tree, we can see that many cell divisions are symmetric and some subtrees occur in more than one place in the tree. We can now perform an assignment of factors such that cells that have identical subtrees also have the same factor content (see Fig. 6). This enables a large simplification, in that it reduces the number of factors and regulators needed to model a given tree. The list of regulators is reduced because all the instances of one subtree are generated by the same set of regulators. Another reduction occurs for symmetric cell division, as described earlier. In this way, the number of regulators is decreased, from 21 needed for the assignment in Fig. 2 above, to 5

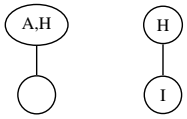

Figure 7: The regulator $g(\mathrm{I}|-\mathrm{A},\mathrm{H})$ denotes that I is not produced if A and H are both present but only if H is present and A is not.

(division) plus 8 (content), in total 13:

$$g(\mathrm{a},\mathrm{b}|\mathrm{A}), g(\mathrm{B}|\mathrm{A},\mathrm{a}), g(\mathrm{C}|\mathrm{A},\mathrm{b}), g(\mathrm{a},\mathrm{b}|\mathrm{B}), g(\mathrm{D}|\mathrm{B},\mathrm{a}),$$
$$g(\mathrm{F}|\mathrm{B},\mathrm{b}), g(\mathrm{a},\mathrm{b}|\mathrm{C}), g(\mathrm{E}|\mathrm{C}), g(\mathrm{a},\mathrm{b}|\mathrm{D}), g(\mathrm{E}|\mathrm{D},\mathrm{a}),$$
$$g(\mathrm{F}|\mathrm{D},\mathrm{b}), g(\mathrm{a},\mathrm{b}|\mathrm{E}), g(\mathrm{F}|\mathrm{E}).$$

The size of this reduced list can be used as a measure of complexity of the lineage (Azevedo et al., 2005; see also the notion of "algorithmic complexity" in Geard and Wiles, 2008). However, further reductions are possible, as will be shown in what follows.

The simplified regulator list above includes several ways to produce factor F: from the expressions $g(\mathrm{F}|\mathrm{E})$, $g(\mathrm{F}|\mathrm{D},\mathrm{b})$ and $g(\mathrm{F}|\mathrm{B},\mathrm{b})$, i.e., the factor F is produced if E is present, or if D or B are present, and then only in the right-hand daughter. The structure is that of a "logical OR", which can be added to the "logical AND" obtained in the discussion on synergy above. The only thing missing to enable full Boolean logic is negation, i.e., "logical NOT."

However, it is acceptable to also include "logical NOT" in the form of repressors within the model, since these are commonplace in gene regulation in biological systems. Transcriptional repressors, in contrast to transcriptional activators, reduce expression of a given gene by binding to DNA sequences in the regulatory regions of the gene. It is important to note that some repressors are not constitutively acting as repressors, but can in some situations switch to become activators, depending on for example the presence of auxiliary regulatory proteins or the chromatin status of the regulatory region (Taatjes et al., 2004). This motivates us to add a general notation to indicate the behaviour of each factor; here, a minus sign. A regulator expression that corresponds to "I is produced if H is present and A is not present" would be denoted (see Fig. 7)

$$g(\mathrm{I}|-\mathrm{A},\mathrm{H}).$$

The mathematical theorem described previously shows that it is possible to represent all trees within the model even without invoking repressors. This may be surprising, since repressors seem to be used in many places in biological systems (Taatjes et al., 2004). One possible reason for the presence of repressors in biological systems is that this may enable a reduction of the required number of factors, or regulator expressions. Including repressors in the model is therefore well motivated, even if not strictly needed to generate general cell lineage trees.

2.5. Modelling complicated factor content

In biological systems, it is almost never only one factor that is responsible for the division and differentiation of cells. The simple procedure of associating one factor to each cell in the tree described above is of course a highly concentrated description in which one specific factor corresponds to a combination of real physical factors, or even different concentrations of the same factor.

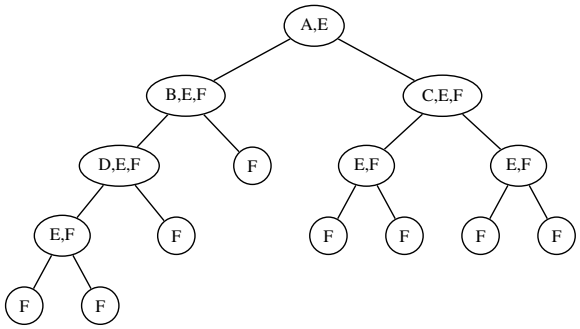

Figure 8: Another assignment of factors that is more economical than previous assignments in terms of regulator expressions.

For example, the trees in Fig. 6 and Fig. 8 are identical in form and end-content, but coded differently; the corresponding regulator list is

$$g(\mathrm{B}|\mathrm{A},\mathrm{a}), g(\mathrm{C}|\mathrm{A},\mathrm{b}), g(\mathrm{D}|\mathrm{B},\mathrm{a}), g(\mathrm{E}|\mathrm{A}), g(\mathrm{E}|\mathrm{B},\mathrm{a}),$$
$$g(\mathrm{E}|\mathrm{C}), g(\mathrm{E}|\mathrm{D},\mathrm{a}), g(\mathrm{F}|\mathrm{E}), g(\mathrm{a},\mathrm{b}|\mathrm{E}),$$

that is, one for division and eight for content, in total nine regulators. The assignment in Fig. 8 uses the explicit choice of letting the factor E control cell division, and also letting it generate the content of the leaves F, as is evident in the regulator list.

This simplifies regulation of cell division from five regulators to one, and generation of F from three regulators to one, but complicates regulation of the factor E from two to four regulator expressions. These changes balance so that the number of content regulators stay the same, even though those for division decrease considerably.

In the above assignment we have moved the OR part of the regulation upward in the tree (less F regulators but more E regulators). Continuing in the same manner, we can simplify the OR part by letting the nodes in question contain more factors. We are effectively moving the OR part to the content of the top node, rather than having it in the regulator list. The result is shown in Fig. 9, and the needed regulator list is

$$g(\mathrm{B}|\mathrm{A},\mathrm{a}), g(\mathrm{C}|\mathrm{A},\mathrm{b}), g(\mathrm{D}|\mathrm{B},\mathrm{a}), g(\mathrm{E}|\mathrm{C}),$$
$$g(\mathrm{E}|\mathrm{D},\mathrm{a}), g(\mathrm{F}|\mathrm{E}), g(\mathrm{a},\mathrm{b}|\mathrm{E});$$

six for content and one for division, in total seven. One OR remains, because of the continuing asymmetric cell division present in the tree.

Now, evidently, the factor A is not needed anymore if we allow for synergy, since the only place where B and C are present simultaneously is when A is present. If we substitute A for the pair [B,C] in the above regulator list,

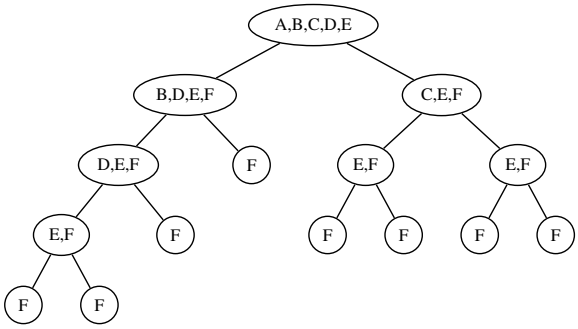

Figure 9: An even simpler regulator list arises from another factor assignment.

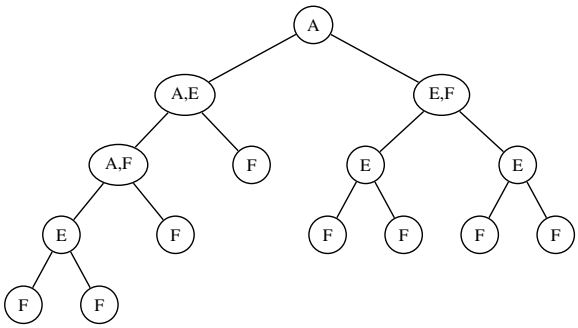

Figure 10: A fully binary assignment is the most economical as regards the number of factors.

we can remove the factor A entirely. This means that reassigning factors in this manner, we have reduced the required factors by one. Indeed, we have simplified both the regulator list and the required factors by assigning a combination of factors to each node, instead of as before, using one factor in each node.

Here, the question arises what representation would use the least amount of factors, using synergy and the elements from Boolean logic we have established. The answer is a fully binary representation where each internal state from Fig. 6 is assigned a binary combination of factors (see Fig. 10). This representation would use the regulators

$$
\begin{aligned}
&g(\mathrm{a},\mathrm{b}|\mathrm{A},-\mathrm{E},-\mathrm{F}),\\
&\quad g(\mathrm{A}|\mathrm{A},-\mathrm{E},-\mathrm{F},\mathrm{a}), g(\mathrm{E}|\mathrm{A},-\mathrm{E},-\mathrm{F},\mathrm{a}),\\
&\quad g(\mathrm{E}|\mathrm{A},-\mathrm{E},-\mathrm{F},\mathrm{b}), g(\mathrm{F}|\mathrm{A},-\mathrm{E},-\mathrm{F},\mathrm{b}),\\
&g(\mathrm{a},\mathrm{b}|\mathrm{A},\mathrm{E},-\mathrm{F}),\\
&\quad g(\mathrm{A}|\mathrm{A},\mathrm{E},-\mathrm{F},\mathrm{a}), g(\mathrm{F}|\mathrm{A},\mathrm{E},-\mathrm{F},\mathrm{a}),\\
&\quad g(\mathrm{F}|\mathrm{A},\mathrm{E},-\mathrm{F},\mathrm{b}),\\
&g(\mathrm{a},\mathrm{b}|\mathrm{A},\mathrm{E},\mathrm{F},\mathrm{b}),\\
&\quad \ldots
\end{aligned}
$$

In terms of *factors* a minimal representation has a number of factors that is the number of bits needed to binary represent the node-associated factors; formally,

$$\log_2(\#\text{Unique subtrees}) \leq \#\text{Factors in binary model.}$$

In our example, there are six unique subtrees, so the minimum number of factors is three. The assignment of Fig. 10 is such that the above regulator list can be reduced greatly, but this is only due to the assignment; there is no guarantee that the regulator list is small for a binary representation. Furthermore, when this reduction can not be made, each regulator contains all the factors, either as activators or repressors. Biological systems do not, as a rule, have this property. While this is the absolute minimum, there is therefore no reason to expect that a biologically interesting system actually operates in this manner.

It is interesting to note that synergistic combinations of factors in the nodes can be used (and will be used below) to reduce the number of factors and regulator expressions that are needed for each tree. There are benefits from the perspective of this model in using synergy, in the form of fewer factors and regulators being required, at the expense of a slight increase in regulator complexity; benefits that can be expected to carry over into the biological system.

2.6. Extending to discrete factor concentrations

Returning to the assignment in Fig. 9, the factor content and its reduction along each branch of the tree seems to imply that the factor content corresponds to decreasing concentrations of one and the same factor, rather than many different factors disappearing one by one. In this language, the combination [C,E,F] corresponds to a high concentration of F; and that the combination [E,F] corresponds to a medium concentration of F, just enough to initiate cell division; while F itself corresponds to a low concentration that does not initiate cell division.

Thus, we propose to represent different concentrations of factors in our model; this is done by using the notation F^3 corresponding to the concentration 3 of the factor F. The unit of concentration can be chosen freely, and this will be discussed in more detail below. To make our new notation include the old notation, we write F for F^1. The regulator expressions denote the change in concentration of each factor, and the changes are added, in contrast to the binary model which only represents if a factor is present or not, and the output of more than one regulator expressions that result in the same product are simply OR-ed together (added, modulo 2).

The former binary model also had a life-time of each binary factor of one time step; the corresponding behaviour here will be to let the concentration decrease by one for each time-step. We therefore include, for each factor, the implicit regulator

$$g(\mathrm{F}^{-1}),$$

which is always active. The concentration levels are bounded below by zero, no negative concentrations are allowed. The specific steps are as follows: the model checks what regulator functions are active, decreases the level of each factor in the cell by one step, and then changes the concentration as instructed by the active regulator functions.

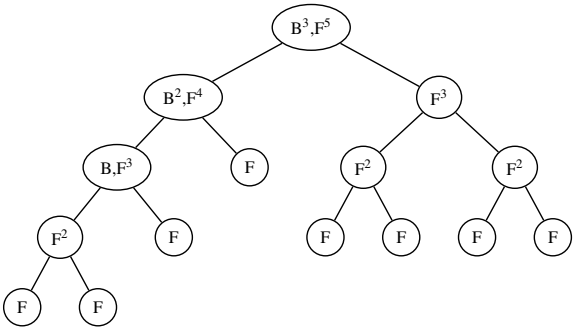

Figure 11: A factor assignment with two factors and decreasing concentration levels.

Concentration changes of factors may also impinge on asymmetric cell division. This can be achieved by specifying additional reduction of the concentration in one of the branches, for example, the regulator $g(\mathrm{F}^{-1}|\mathrm{b})$ would make the factor decrease another concentration-level in the right-hand daughter of each division, in addition to the implicit decrease mentioned above. Note that this regulator in general does not correspond to transcriptional changes but rather to processes such as asymmetric distribution.

One assignment that utilizes this fully is presented in Fig. 11, where there are two factors: one that controls cell division and another that makes the tree have the appropriate asymmetry. The four regulator expressions needed and their meaning are as follows:

$g(\mathrm{a}, \mathrm{b}|\mathrm{F}^2)$: Cell division is initiated if the concentration of F is two or more.

$g(\mathrm{F}^{-1}|\mathrm{B}, \mathrm{b})$: The concentration of F decreases by one per time step (implicitly), but is asymmetrically distributed if B is present.

$g(\mathrm{F}^{-1}|\mathrm{B}^2, -\mathrm{F}^5, \mathrm{b})$: The concentration of F is even more asymmetric if the concentration of B is high, unless the concentration of F is very high.

$g(\mathrm{B}^{-2}|\mathrm{b})$: The concentration of B decreases by one per time step (implicitly) and B is completely asymmetric in that it always goes to the left-hand daughter.

An important observation is that the assignment in Fig. 11 does not rely on transcription, but only on degradation to reduce the concentration. All factors are present in the root, and only decreases as time proceeds, so this asymmetric tree is possible to create without transcription. This may be similar to the very earliest stages of differentiation, where the first cell divisions rely on maternal mRNA laid down in the egg, and thus can proceed without de novo mRNA synthesis. When comparing Fig. 6 and Fig. 11, there are obvious correspondences:

A corresponds to $[\mathrm{B}^3, \mathrm{F}^5]$;
B corresponds to $[\mathrm{B}^2, \mathrm{F}^4]$;
C corresponds to $[\mathrm{F}^3]$;
D corresponds to $[\mathrm{B}, \mathrm{F}^3]$; and
E corresponds to $[\mathrm{F}^2]$.

In other words, the factors from Fig. 6 are represented as combinations of the two factors at different concentrations in Fig. 11. The regulator list is simpler, and the list for Fig. 6 seems to create new factors in each step, (and thus seems to need transcription), all the while the assignment in Fig. 11 does not.

It may seem as if the assignment in Fig. 11 reaches below the minimum of three factors needed to model the tree, since only two factors are used. This is however not the case, since we have moved from the binary model to a model that includes discrete concentration levels. The possible states in this model are actually many more than the required six in this case, since we have two factors that can have five different concentrations each. There are restrictions as we want the assignment to be plausible from a biological point of view, but if these are ignored only one factor at six different levels of concentration is needed. The details will not be given here since the regulators in this case do not have the simple decreasing structure as in the previous example, and are not plausible from the biological point of view.

As has been discussed earlier, the factors in this model may or may not correspond to physical proteins. The factors in this model should be thought of as "meta-factors," (hence the name meta-Boolean) and a model factor can correspond to a combination of several physical factors, a physical factor at some specified concentration, external influence such as intracellular signalling, or indeed, even external pressure. The representation and modelling of external influence will be further discussed below.

As a final remark, it is immediately obvious that even for the small example tree we have used here, there is a very large set of possible factor assignments and regulator lists. There are even several regulator lists for each factor assignment. Some of these lists are more plausible from the biological point of view than others. It is therefore not likely that a random search through assignments and regulator lists would yield a biologically relevant model, or one that is better in that respect than a factor assignment simply chosen according to some rule. In what follows, the explicit assignment procedure and reduction of Figs. 2 and 6 will be used, together with ideas for reduction of factor creation along the line as that used in Fig. 11.

3. Testing the model on the embryonic *C. elegans* cell lineage tree

Up to this point, the discussion has been on generic properties of the model, and it will thus be important to validate the model on a true cell lineage tree, with various differentiated cell types as terminal nodes in the tree. We will demonstrate usage of the model in the *C. elegans* embryonic cell lineage tree, and therefore we first describe some key aspects of the tree.

3.1. Structure of the C. elegans *cell lineage tree*

As our example, we will use the *C. elegans* hermaphrodite cell lineage as specified in Sulston et al. (1983). This lineage tree contains 1341 cells of which 671 are terminal cells, whereas 670 cells are found in intermediate positions in the lineage. The terminal cells are post-mitotic, fully differentiated cells, and many of these cells can be can be grouped into a few distinct categories with regard to cell type (muscle, neuron, intestine, hypodermis, ...). Some of the terminal cells will be eliminated by programmed cell death, and we count these as a distinct cell type below, despite the fact that they eventually die. For the purposes of this example, to demonstrate the presented model, a coarse-grained grouping of the cell types will be sufficient. A more detailed tree can equally well be described by the model, but would be unnecessarily complicated here.

At intermediate positions in the lineage, cells can be considered identical if they have the same ancestral pedigree and generate the same type of daughter cells. In some parts of the cell lineage, there are "sub trees" in the lineage, which are highly related, for example see the muscle part of the "C" subtree as shown in Fig. 12. Among the terminal cells of the tree are also 58 cells that will temporally stop dividing in the embryonic cell lineage, but will again assume cell division at later stages, and these can be treated much as the intermediate cells: they belong to the same group if they eventually generate the same type of daughter cells.

Of the 670 intermediate cells in the cell lineage tree, 483 cells seem to undergo asymmetric cell division in that they generate distinct daughter cells. But it is known that some of the cell divisions that are asymmetric *in the tree* (as regards the cell types of the daughters) actually are internally symmetric cell divisions augmented by external influence from neighbouring cells. The earliest example of this in *C. elegans* is in the second cell division, where the lineage tree is asymmetric in both divisions, but only the division in the branch that eventually gives the germ line is asymmetric (Platzer and Meinzer, 2004, and references therein).

3.2. An example: the "C" subtree of the C. elegans *lineage tree*

The previously shown *C. elegans* lineage subtree (Fig. 5) was modelled using an assignment with node-specific factors, with the exception of the leaves, and this technique can be used to model the entire *C. elegans* cell lineage tree (as that present at `www.wormatlas.org`). The resulting list of regulator expressions can be found as supplementary material accompanying this paper. However, a model with node-specific factors is not very useful because the content of each node is hidden within the node-specific factor, perhaps representing a combination of physical factors that give the properties of that particular node. It is more interesting to look at attempts to reduce the number of factors, the number of regulator expressions, and above all, attempts to follow the biological behaviour. We can have a closer look at this process using a larger portion of the *C. elegans* lineage tree than in Fig. 5, but not the entire tree; we will look at the part commonly labelled the "C" part, see Fig. 12.

The large number of states in Fig. 12 can be reduced by identifying the cells that have identical subtrees. This makes for a reduction from the 51 node-specific factors of Fig. 12 to a total number of 18, see Fig. 13.

To reduce the factors even further, let us start by introducing a factor C to control the cell division, much as in the previous example. We want the cell division to continue for five generations, so the initial amount of the factor should be 5, as in Fig. 14a. This tree only needs the regulator expression

$$g(\mathrm{a}, \mathrm{b}|\mathrm{C}).$$

It is also possible to let the factor H be created at the high initial level of C, making all daughters eventually differentiate into hypodermis cells, as in Fig. 14b. To achieve this, we only need to add the regulator expression

$$g(\mathrm{H}^5|\mathrm{C}^5).$$

This suffices for producing a completely symmetric tree that ends in hypodermis cells; indeed, it would so far be possible to make due with only one factor, and letting a high level of this factor control cell division. However, the needed differentiation is easier to achieve within the model if differentiation is controlled by a separate factor from that which controls cell division. Note that the levels of the factor H may be interpreted, not as a high concentration that decreases to a specified level, but instead a low concentration that increases to a stable level (see Section 5.1), perhaps controlled by some self-regulatory transcription process not included in the present description.

Now, to create the muscle part of this subtree, we let the highest level of the factor H produce a high level of the factor M in the right-hand daughter (see Fig. 14c). To do this, we need to add the regulator expression

$$g(\mathrm{M}^5|\mathrm{H}^5, \mathrm{b}).$$

We now need an additional cell division in the last step of the muscle part of the tree. There are several options here, one would be to simply add one level of C at the creation of the factor M above, or rather, hinder the decrease

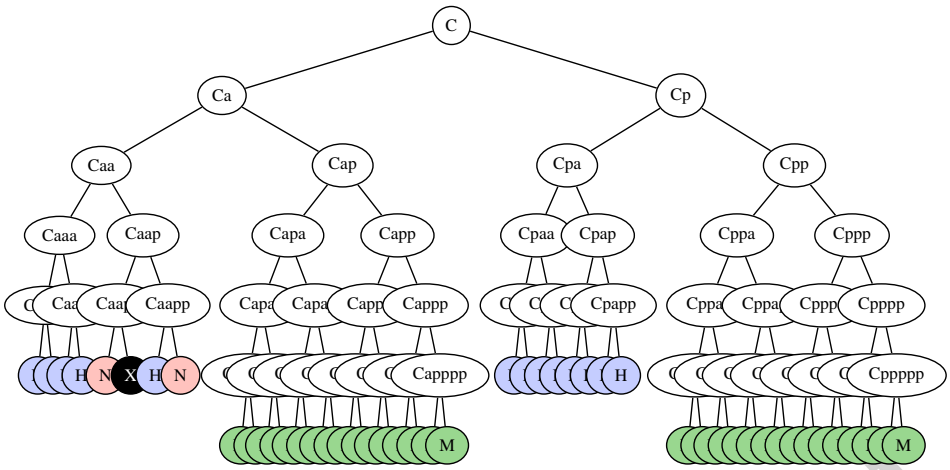

Figure 12: The "C" part of the *C. elegans* cell lineage tree, using node-specific factors, except for the leaves. There are 4 terminal cell types: hypodermis (H), nerve (N), programmed cell death (X), and muscle (M). This factor assignment requires 51 factors, 79 factor-specific regulator expressions and 47 division regulator expressions.

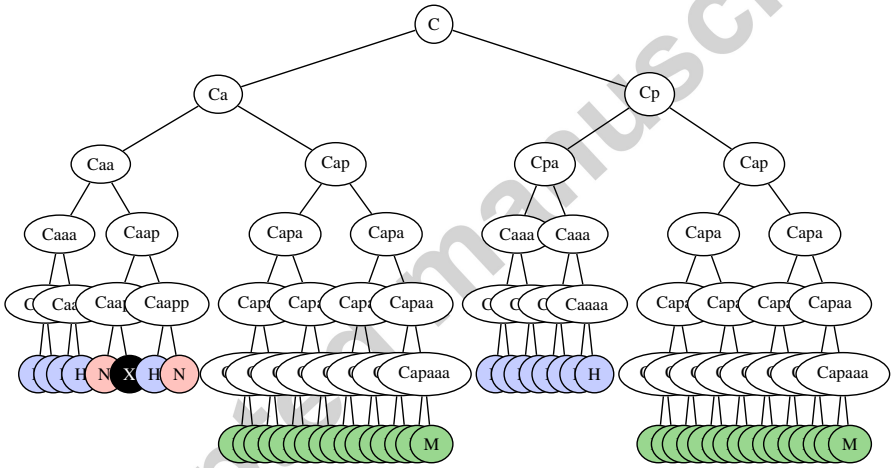

Figure 13: The "C" part of the *C. elegans* cell lineage tree, using node-specific factors and reduction by identification of equal subtrees. For example, the two muscle subtrees (ending in M) are identical in differentiation and form and therefore, the assignment here is made so that both subtrees start with Cap, instead of as in Fig. 12, starting with Cap and Cpp, resp.. The 4 terminal cell types remain, and the factor assignment requires in total 18 factors, 21 factor-specific regulator expressions and 14 division regulator expressions.

in the step from mother cell to right-hand daughter in that step. But this makes the cell-division factor C behave asymmetrically in one step of the process, which seems like an unnecessary complication. A simpler, more biologically plausible way to do this would be to let the factor M slow the rate of decrease of C in the cells. This would mean that the level of C would decrease in non-integer steps (see Fig. 15) by using the regulator expression

$g(\mathrm{C}^{0.4}|\mathrm{M})$: C^n decreases to $\mathrm{C}^{n-0.6}$ instead of C^{n-1} if M is present.

Although it may appear that the factor M causes transcription of the factor C, this need not be the case. It may simply be that the deterioration of the factor C is slower when the factor M is present, or the cause may be that the cell division happens more rapidly when the factor M is present; that the time-steps in our model happens more often in real time when M is present.

The slowed decrease of C in the muscle subtree ensures that the division continues for one more time step. In the above example, three time steps need to become four, which means the 3 units of C present at the top node of the muscle subtree cannot decrease too fast; there needs to remain at least 1 unit after three time steps to give another division. This means that the rate of decrease needs to be less than 2/3 since a larger decrease would lead to a too rapid decline in concentration. At the same time, the decrease needs to be strictly larger than 1/2, since a smaller decrease would make the divisions continue for too long. The choice 0.6 is somewhat arbitrary (used here because the decimal expansion is short), but as mentioned above, even having a slowed decrease on this form is only

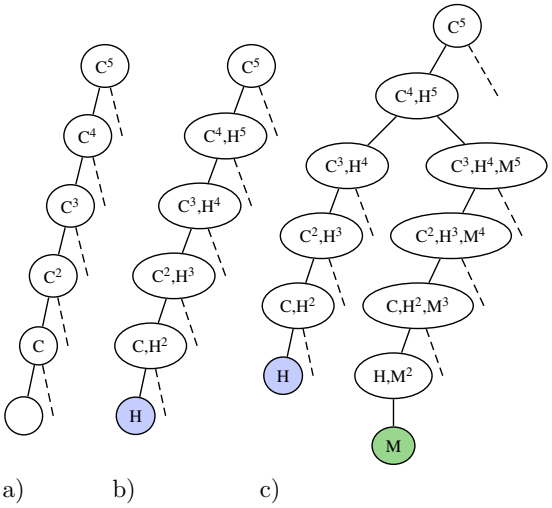

Figure 14: Further reducing the representation by a) letting the factor C control cell division and start at a high level, b) letting the factor H be created at this high level of C, and c) letting the factor M be created in the right-hand daughter at a high level of H. Repeated cells are suppressed in the above images.

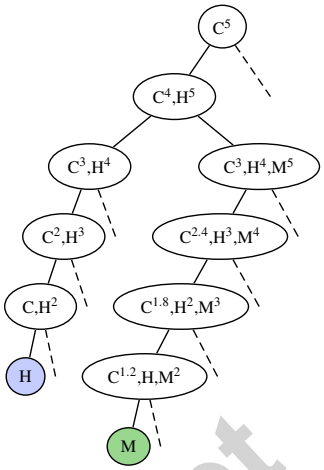

Figure 15: Letting the factor M slow the decrease of the cell division controlling factor C, note the extra cell division. The factor $C^{0.6}$ is suppressed from the terminal muscle cells, because we have implicitly assumed that the level 1 is the lowest level where any factor contributes to the system in any way; level 1 is the threshold below which the factor has no effect.

one of many choices. This is because of the large set of possible factor assignments and regulator lists that gives the same tree, as discussed at the end of Section 2.

Using non-integer levels may seem like a step towards a level-continuous model, but this is not the case. What we have done here is to divide the integer levels of C fivefold. The usual integer levels of C can now be thought of as 5/5 (five fifths) of C, and the added decrease 0.6 can be thought of as 3/5 of C, so that the levels now are an integer number of fifths of C rather than in whole units of C. In a general mathematical tree (that could be infinite), this could lead to infinite subdivision of concentration levels, or in other words, to a level-continuous model. This model is intended specifically for finite-length and usually relatively short cell lineage trees, and then, there is no such danger. For this application, the model is still level-discrete.

Now, for the asymmetry of the "C" subtree; the presence of nerve cells and programmed cell death in the left half of the tree. This calls for an asymmetric cell division in the first generation using, say, the nerve cell factor N, using the regulator function

$$g(\mathrm{N}|\mathrm{C}^5, \mathrm{a})$$

Let us continue this with an increase in N in the left-hand daughter if N is present and the level of C is high enough

$$g(\mathrm{N}^{1.4}|\mathrm{C}^4, \mathrm{N}, \mathrm{a}),$$

going on with a slower increase in the right-hand daughters if the level of N is high enough and C is present at a lower level,

$$g(\mathrm{N}^{1.2}|\mathrm{C}, \mathrm{N}^{1.4}, \mathrm{b}).$$

Finally, at a certain point, we need the left-hand daughter to increase the N content, making the terminal left-hand daughter a nerve cell. It is actually easiest to let the regulator be active at the C^2 level, because then the result can be used to control the programmed cell death as well through

$$g(\mathrm{N}^{2.4}|\mathrm{C}^2, \mathrm{N}^{1.6}, \mathrm{a}), g(\text{cell death}|\mathrm{N}^3, \mathrm{b}).$$

The expressions involving N are, similarly as above, not very sensitive to the actual values of the numeric expressions, although they do seem somewhat arbitrary (more on this in Section 4 below). The resulting tree can be seen in Fig. 16.

There are several alternatives here, just as in the previous simple example. For example, the regulator function involving C^4 can be made symmetric at the price of introducing M as a repressor in the later regulator functions so that the factor N disappears from the muscle subtree after the first few generations. This would make the muscle subtrees unequal since the factor N would be present at the top of one of them. This shows that it is not strictly necessary to have identical subtrees as hinted in Fig. 13, but it is possible, and perhaps simpler from one point of view to have unequal subtrees even if they are equal in form. On the other hand, it can be undesirable to have the factor N in the muscle subtree; but we leave this question for now. It would be fair to suspect that there is an almost endless supply of alternatives of this kind, even for this relatively small part of a lineage tree.

4. Cell-extrinsic factors

In addition to only exploring consequences of altering cell-intrinsic parameters, it is of interest to also address

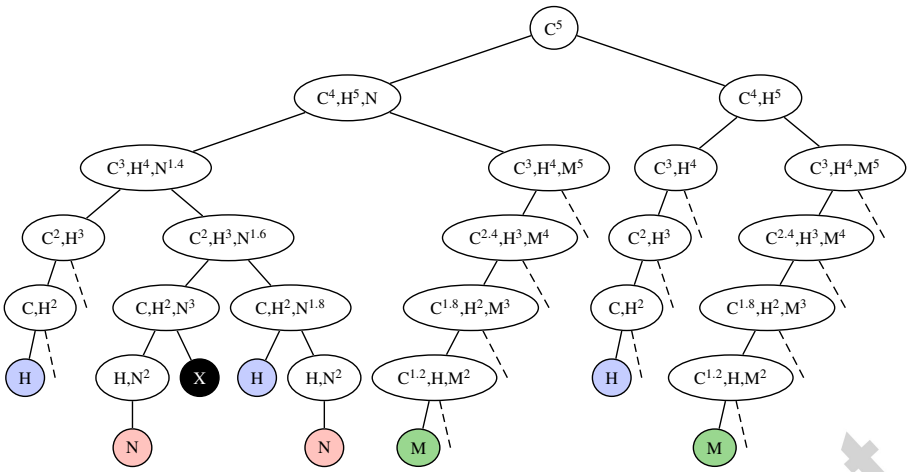

Figure 16: The complete "C" subtree, using five factors and nine regulator expressions. Identical branches are suppressed.

how the inclusion of cell-extrinsic factors affects the complexity of the models. Cell-extrinsic factors play a role in communication between cells and can act on longer distances, for example in the form of secreted hormones or growth factors, but also between cells in direct contact, for example by the Notch signalling mechanism. We will now have a brief look at cell-extrinsic regulation and how it can be included in the model. Thus far, if one needs asymmetry in the lineage tree, the tree needs to be asymmetric at the top, in the first division. This is evident in the example above, where the cell death and nerve cells need an asymmetry already in the very first division. A cell-extrinsic influence, caused for example by another cell in the lineage, would make it possible to obtain asymmetry without having to resort to asymmetric cell division. Evidence for the importance of cell-cell interaction comes from many parts of the *C. elegans* embryonic lineage, as mentioned one example can be obtained already in the second cell division (Platzer and Meinzer, 2004).

In the "C" subtree example above, an asymmetry is introduced already at the first division to enable differentiation to nerve cells in only the left-hand part of the tree. Deleting the asymmetry in the top division, and the subsequent regulator functions that involve the nerve factor N will return us to the tree in Fig. 14. At this point, we have the regulator list

$$g(\mathrm{a},\mathrm{b}|\mathrm{C}), g(\mathrm{H}^5|\mathrm{C}^5), g(\mathrm{M}^5|\mathrm{H}^5,\mathrm{b}), g(\mathrm{C}^{0.4}|\mathrm{M}).$$

To return the tree to the desired form, it would be possible to add (identical) external influence on the two cells indicated in Fig. 17, and also add the regulators

$$g(\mathrm{N}^2|\text{ext. infl.}), g(\text{cell death}|\text{ext. infl.},\mathrm{b}).$$

There are now four different terminal cell types (including programmed cell death), each represented with their own factor, one factor that regulates cell division, and one that represents the external influence; in total six factors. There are also six regulator expressions: one for each internal factor, and one that controls cell division. We believe that this model captures the desired behaviour in a natural way, and is as economical as possible given the need to control cell division and differentiation into four terminal cell types, and given the asymmetric structure of the tree.

Incorporating external influence as an element in this model immediately decreases the complexity in the regulator expressions. It also simplifies the cell content significantly, and alleviates the need for a number of asymmetric divisions, most prominently at the root of the tree. This suggests that external influence is not only possible, but necessary to describe a cell lineage tree in a factor- and regulator-conservative manner. As mentioned above, it is already known that cell-cell signalling plays a role at the second division in *C. elegans*, and suspected to be important in subsequent divisions.

We would suggest that the presented model will serve as a tool to find places in the *C. elegans* lineage tree where this external influence is important, where the decrease in complexity of the model is large. That is, to predict (or suggest) interesting regions of the embryo where cell-cell signalling takes place. Another use of the present model would be to test known extrinsic factors and the simplifications they enable, as in the C subtree above where the external influence simplifies the model greatly.

The present model can also be used as a platform for integration of data derived from models of signaling crosstalk and information about spatial distribution of cells in *C. elegans*, to make more definitive hypotheses about the extent and location of cell-extrinsic effects on the lineage. Recently, a detailed spatio-temporal description of the positions of cells, and which cells that are engaged in direct cell-cell contacts has also been published (Hench et al., 2009). A cursory look at the supplemental data provided by Hench et al. (2009), although inconclusive at more than 150 cells in the lineage, suggests that one possible external influence as indicated in Fig. 17, could be contact be-

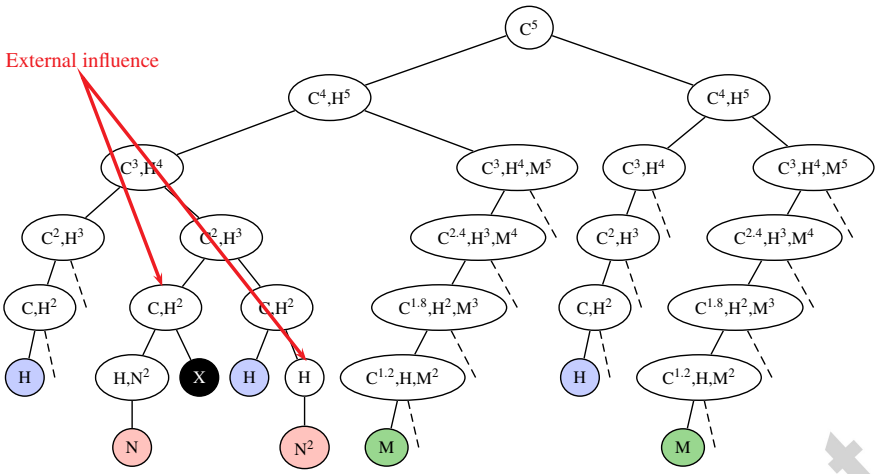

Figure 17: The complete "C" subtree, using six factors (including one for external influence) and six regulator expressions. Identical branches are suppressed.

tween the cells Caapa, Caapp and ABplapapap. In contrast, there appears to be no or very brief contacts between the ABplapapap cell and other cells in the C subtree at this level, although this needs to be further analyzed. This approach can now be used to extend studies like the one in Fisher et al. (2007) where the intricate intersection between two cell-extrinsic signaling mechanisms was analyzed (EGF signaling and Notch/Lin-12 signaling), revealing novel feedback loops. By combining information from the model presented here, signaling cross-talk models and data on the precise spatio-temporal (4D) localization of cells, refined hypotheses for cell-cell interaction can be formulated and directly tested by various experimental techniques, including laser ablation of specific cells, targeted expression of RNAi to remove gene function in specific cells or by analysis of specific gene knock out phenotypes.

5. Comparison with level-continuous, time-discrete models

It is sometimes stated that a Boolean model has less descriptive power than a continuous model (see, e.g., Plahte et al., 1994; Smolen et al., 2000). For the purposes here, this is not the case; we have already noted that the Boolean model can generate any lineage tree, even without using synergy, repressors, or external influence. Actually, the model presented here has the advantage of being simple to handle, and we have shown that synergy, repressors, and external influence can be used to decrease the complexity of the model.

5.1. Discrete concentration levels

Morphogenic activity is often present in biological systems, so that one factor can induce several cell fates depending on if the concentration is above or below certain thresholds. Our model naturally includes this behaviour

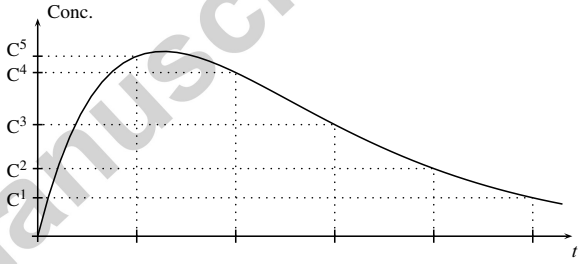

Figure 18: Example of quantization of factor concentration at (here evenly spaced) points in time, in a system with production of the factor C followed by degradation.

by the quantization used. Indeed, in our model, a quantization of concentration levels is directly invoked by the natural discretization of the time axis given by cell division. In each time frame, any single factor has a well-defined range of concentration, and thus, natural quantization to use are for example the highest or lowest point in the interval, the mean over the interval, or simply a number that signifies the number of quantized concentrations that are lower than the present concentration (see for example Fig. 18).

Even rapid fluctuation of the concentration can be handled with this "labelling" of the factor content. Indeed, it is not limited to the case of a clearly defined level of concentration, but can equally well be used to model frequency dependence as in intracellular signalling where different signals are transmitted using the same chemical substance, but where the frequency of the concentration changes carry the information.

We note that quantization *is* a drawback in models where a continuous time coordinate is used, or for example when analyzing stability of stationary points in a dynamic GRN. It has been shown that Boolean (or discrete-level) models are less suitable for that kind of analysis (e.g., in

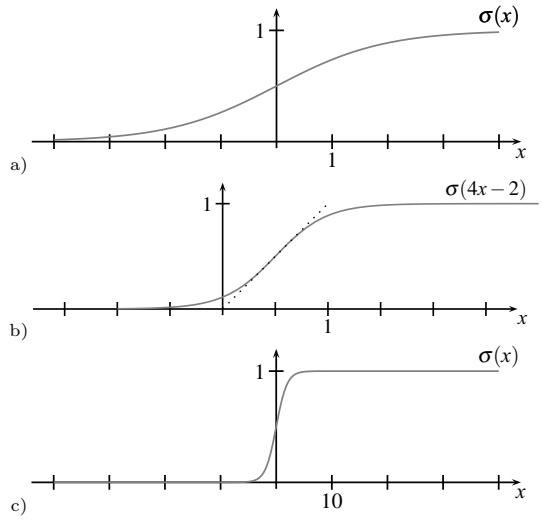

Figure 19: A sigmoid function, normally used in a DRGN to limit the gene expression to the region between 0 (not expressed) and 1 (fully expressed). a) The sigmoid function used here is the commonly chosen "logistic curve" $\sigma(x) = 1/(1+\exp(-x))$. b) The same sigmoid function, scaled and shifted. This function is almost $y = x$ between 0 and 1 (dotted line) and limited between 0 and 1. c) The same sigmoid function with a larger portion of the x-axis visible.

Plahte et al., 1994, but see also Smolen et al., 2000), but in the case of analyzing the structure of a lineage tree the *changes* in factor content is the important feature. Therefore, the present model can be used for this purpose without loss of generality, allowing for simple modelling of generic cell lineage trees, including the described reduction of used factors by using identification of subtrees, synergy effects and external influence.

5.2. Synergy in a Dynamic Recurrent Gene Network

Another question is if the explicitly introduced synergy effects in the presented model prevents comparison with the usual linear framework in some GRN modelling attempts. In a general GRN, there are usually a number of synergic influences, but many models for a GRN seem to prohibit this kind of effects because of the linearity of mathematical functions used. For example, a Dynamic Recurrent Gene Network (DRGN) as that used in Geard and Wiles (2005) and Lohaus et al. (2007) uses a linear map from one time step to the next. However, there is a nonlinear function that limits the activity of the genes in each time step see Fig. 19. We will briefly show here that this nonlinearity makes synergy effects possible, and in fact enables full Boolean logic within the model.

To see this, we first need some notation: within the DRGN, we label the expression of the ith gene at time t by $x_i(t)$ (similarly as in Geard and Wiles, 2005). The level of expression for gene i at a time t is given by a linear function of the level of expression of all genes at the previous time step $x_1(t-1)$, $x_2(t-1)$, ..., limited between 0 and 1 by a sigmoid function. As an example, the expression of gene 2 can be influenced by the expression of gene 1 in the previous time step via

$$x_2(t) = \sigma\Big(4x_1(t-1) - 2\Big).$$

This would yield a value of $x_2(t)$ that is almost equal to $x_1(t-1)$ in the interval zero to one (i.e., almost linear), but is limited between 0 and 1, see Fig. 19b.

Now, the sigmoid function constitutes an explicit nonlinearity in the model. While this function sometimes is described as merely restricting the gene expression to the interval between 0 (not expressed) and 1 (fully expressed), it was observed already in Walter et al. (1967) that a sigmoid limitation like this would enable on-off switching, one basic trait of binary logic. This is evident if we look at a larger interval of the x-axis, see Fig. 19c.

Using this observation, it is simple to model on-off switching, for example via

$$x_3(t) = \sigma\Big(20x_1(t-1) - 10\Big),$$

because this means that gene 3 is expressed if gene 1 was expressed in the previous time step. Furthermore, the OR of Boolean logic can be represented here via

$$x_4(t) = \sigma\Big(20x_1(t-1) + 20x_2(t-1) - 10\Big),$$

where gene 4 is expressed if gene 1 or gene 2 (or both) were expressed in the previous time step. This might not be surprising since OR is just a Boolean representation of addition (XOR is addition modulo 2, OR is upper-bounded addition). However, the AND of Boolean logic, corresponding to multiplication (modulo 2), can also be represented via

$$x_5(t) = \sigma\Big(20x_1(t-1) + 20x_2(t-1) - 30\Big).$$

That this corresponds to logical AND can be verified visually in Fig. 19c; gene 5 will be expressed if *both* gene 1 and gene 2 were expressed in the previous time step, but not if only one or none of them were. For completeness, repression (logical NOT) has a natural representation, for example

$$x_6(t) = \sigma\Big(-20x_1(t-1) + 10\Big),$$

will give expression of gene 6 if gene 1 was *not* expressed in the previous time step.

Some care needs to be taken, for example when combining an activator and a repressor in an AND expression, but these are simple to handle. There is, however, one specific complication that needs to be noted here: the combination of two ANDs via an OR. This complication is simplest to illustrate via example, so suppose that we want a gene to be expressed if (gene 1 AND gene 2) OR (gene 3 AND gene 4) were expressed in the previous time step. The first AND is already given above, controlling the expression of gene 5, and the second AND would be given by

$$x_7(t) = \sigma\Big(20x_3(t-1) + 20x_4(t-1) - 30\Big).$$

But we cannot write

$$x_8(t) = \sigma\Big(20x_1(t-1) + 20x_2(t-1) + 20x_3(t-1) + 20x_4(t-1) - 30\Big),$$

because that would correspond to expression of gene 8 if *any two* of gene 1, 2, 3, and 4 were expressed in the previous time step. In fact, since the inner expression is linear in the x_js, we can represent an OR expression or an AND expression, but not a combination of the two. To handle this, we need to allow *both* gene 5 and gene 7 to correspond to the same biological gene. This means that it is still possible to represent this kind of synergy effect in a DRGN, but there is a price to pay as an increased number of gene indices i. Indeed, several synergies that activate the same gene would be possible to model in a DRGN but this requires the same meta-language (meta-factors) to be used as in our meta-Boolean model, in this case, several model genes that correspond to the same biological gene. In special situations (e.g., when x_1 and x_2 are zero whenever x_3 and x_4 are nonzero), this can be handled without an increase in model-gene number, but a detailed list of these situations is out of the scope of the present paper.

A last remark here is that the deviation from the ideal 0 and 1 in these expressions is less than 5×10^{-5}, so it would take many time steps (in our case, cell divisions) to produce a sizable deviation. We also note that the deviation decreases exponentially with the size of the factors 10 and 20 used above.

5.3. Creating a DRGN from a lineage tree

Using these expressions, it is possible to directly create a DRGN that gives a lineage tree with the correct structure and terminal cell content. The procedure is simply to create a Boolean model of the lineage tree using node-specific factors as in Section 2.4 and translate that into the DRGN formalism using the maps just mentioned.

As regards asymmetric cell division, for example in Geard and Wiles (2005), the representation is slightly different than here; instead of our factors a and b, asymmetric cell division is handled by a "relative position" input $I(t)$ that is 1 when producing the right-hand daughter and 0 when producing the left-hand daughter. Translating to our notation, the relative position I acts as an activator on the right and a repressor on the left, so that, for example $g(\mathrm{C}|\mathrm{F},\mathrm{b})$ corresponds to an AND expression where the relative position I is an activator for the factor C,

$$x_{\mathrm{C}}(t) = \sigma\Big(20I(t-1) + 20x_{\mathrm{F}}(t-1) - 30\Big)$$

(where the numeric indices of the factors C and F should be substituted). Similarly $g(\mathrm{B}|\mathrm{F},\mathrm{a})$ corresponds to an AND expression where the relative position input I is a repressor for the factor B,

$$x_{\mathrm{B}}(t) = \sigma\Big(-20I(t-1) + 20x_{\mathrm{F}}(t-1) - 10\Big).$$

We note that the node-specific assignment will give a DRGN that generates the correct tree in terms of structure and differentiation, and that this procedure does not need iterative search techniques like the one used in Geard and Wiles (2005). Unfortunately, the produced DRGN will contain many node-specific factors that will hide the biological factor content, and this is not desired. It seems much more fruitful to use the reduced representation from Section 2.6 representing discrete concentration levels of the factors, and from that derive the map to use in a DRGN to get from one time-step to the next. This map is less immediate to produce, but the proposed formalism is promising.

6. Conclusions

We have presented an extension of the classical Boolean model, the "meta-Boolean model", aimed at handling cell lineages. The additions are primarily related to mechanisms that control factor levels, cell differentiation and cell division. We have further discussed how the complexity of the model in terms of number of factors and regulators is influenced by different assumptions such as the use of repressors, AND-OR cis-regulatory logic, multiple levels, the additional mechanism for factor level changes and external factors.

A benefit of the model is that it is simple to handle and that it allows direct description and generation of large cell lineage trees. The model can easily incorporate external influence from other cells or the environment. This will be very beneficial when modelling more complicated processes such as the competitive interaction between cells generated along the mid-line of the *C. elegans* embryo (Geard and Wiles, 2005; Sulston et al., 1983).

It is often stated that Boolean (or discrete-level) models have less expressive power than continuous-level models. This is true when discussing stability and features that depend on low-level description of the biological processes, but when used to study differentiation, the present model is general in the sense that it it can be used to describe any cell lineage tree.

The new model has been applied to the *C. elegans* cell lineage tree. Explicit results concerns the trade-offs between the number of regulators and extrinsic factors, in the C subtree. The model is also capable of giving indications regarding co-variation of factors, number of involved genes and where in the cell lineage tree that asymmetry might be controlled by external influence. We would suggest that the presented model will serve as a tool to find places in cell lineage trees where this external influence is important (e.g., in *C. elegans*), where the decrease in complexity of the model is large. In combination with a spatio-temporal description of cell positions, it can be used to predict (or suggest) interesting regions of the embryo where cell-cell signalling takes place. Adding knowledge on known modes of external signaling would enable direct tests, for example through laser ablation, RNA interference, or usage of knock out phenotypes.

We have furthermore shown that the model is capable to emulate linear differential/difference models capped by a sigmoid function. Even though these latter models may at first seem to be essentially linear, the use of a nonlinear limiting function enables synergy within the models, at the price of a possible increase in the number of model factors. In our Boolean-based model, such synergy is directly included rather than being an effect of a limiting function. Having established a dependence (synergic or not) in the present model, it is simple to translate into the language of a difference-equation model. Thus, the model can also be used as a tool to design more complex difference-equation models.

The proposed model is quite general and more work is needed to find biologically plausible constraints while still retaining the generality. In particular, the factors and regulators have only been addressed formally in this paper, and some effort is still needed to match these theoretical constructs to known biological counterparts.

In conclusion, we believe that the model and the associated terminology presented here is a powerful tool and can serve as a platform for more detailed mappings of the biological factors that control cell lineage creation, synergy and external influence.

Appendix: The model is general

Here we briefly give a mathematical theorem showing that any cell lineage tree can be generated with our model. A cell lineage tree with node-specific labeling is known as a "labeled ordered rooted complete binary tree" (the nodes have unique labels, there is a natural left-right order of the daughter nodes, the tree has a natural starting point, all parent nodes have two daughters). We use the model as described in Section 2.3 restricted to the regulator expressions of Fig. 4a-b. The following result can now be obtained.

Theorem: Any labeled ordered rooted complete binary tree can be generated from the root node by using an appropriate list of regulator expressions.

Proof: We must first establish the list of regulator expressions. For each intermediate node in the tree, add three regulator expressions:

(i) $g(\mathrm{a}, \mathrm{b}|x)$, where x is the label (the node-specific factor) of the current intermediate node;

(ii) $g(y|x, \mathrm{a})$, where y is the label of the left-hand daughter;

(iii) $g(z|x, \mathrm{b})$, where z is the label of the right-hand daughter.

The procedure to generate the tree is now the following:

0. Start with the root node;
1. try to find $g(\mathrm{a}, \mathrm{b}|x)$ in the list of regulator expressions, where x is the label of the current node;
2. if successful,
   (a) then there are expressions $g(y|x, \mathrm{a})$ and $g(z|x, \mathrm{b})$ in the list; use these to generate the daughters and their labels;
   (b) perform steps 1 and 2 for the left-hand daughter;
   (c) perform steps 1 and 2 for the right-hand daughter.

This procedure will recursively go through each intermediate node in the tree and expand it. Each branch will stop only at a terminal node; the procedure will generate the entire tree from the starting node. □

References

References

I. Albert, J. Thakar, S. Li, R. Zhang, and R. Albert. Boolean network simulations for life scientists. *Source Code for Biology and Medicine*, 3, 2008.

R. B. R. Azevedo, R. Lohaus, V. Braun, M. Gumbel, M. Umamaheshwar, P.-M. Agapow, W. Houthoofd, U. Platzer, G. Borgonle, H.-P. Meinzer, and A. M. Lerol. The simplicity of metazoan cell lineages. *Nature*, 433:152–156, 2005.

A. J. Bannister and T. Kouzarides. Differentiation and gene regulation. *Current Opinion in Genetics and Development*, 15(5 SPEC. ISS.):473–475, 2005.

W. Banzhaf. On the dynamics of an artificial regulatory network. In W. Banzhaf, T. Christaller, P. Dittrich, J. T. Kim, and J. Ziegler, editors, *Advances in Artificial Life,* 7th European Conference, ECAL 2003, Dortmund, Germany, September 14-17, volume 2801 of *Lecture Notes in Computer Science*, pages 217–227. Springer, 2003. ISBN 978-3-540-20057-4.

G. Bernot, J.-P. Comet, A. Richard, and J. Guespin. Application of formal methods to biological regulatory networks: Extending Thomas' asynchronous logical approach with temporal logic. *Journal of Theoretical Biology*, 229:339–347, 2004.

J. W. Bodnar. Programming the Drosophila embryo. *Journal of Theoretical Biology*, 188:391–445, 1997.

J. W. Bodnar and M. K. Bradley. Programming the Drosophila embryo 2: From genotype to phenotype. *Cell Biochemistry and Biophysics*, 34:153–190, 2001.

V. Braun, R. B. R. Azevedo, M. Gumbel, P.-M. Agapow, A. M. Leroi, and H.-P. Meinzer. ALES: Cell lineage analysis and mapping of developmental events. *Bioinformatics*, 19:851–858, 2003.

E. H. Davidson. The sea urchin genome: Where will it lead us? *Science*, 314:939–940, 2006.

H. de Jong. Modeling and simulation of genetic regulatory systems: A literature review. *Journal of Computational Biology*, 9:67–103, 2002.

E. Dubrova, M. Teslenko, and A. Martinelli. Kauffman networks: Analysis and applications. In *IEEE/ACM International Conference on Computer-Aided Design, Digest of Technical Papers, ICCAD*, volume 2005, pages 478–483, 2005.

J. Fisher, N. Piterman, A. Hajnal, and T. A. Henzinger. Predictive modeling of signaling crosstalk during c. elegans vulval development. *PLoS Computational Biology*, 3(5):0862–0873, 2007.

C. Furusawa and K. Kaneko. Emergence of multicellular organisms with dynamic differentiation and spatial pattern. *Artificial Life*, 4:79–93, 1998a.

C. Furusawa and K. Kaneko. Emergence of rules in cell society: Differentiation, hierarchy, and stability. *Bulletin of Mathematical Biology*, 60:659–687, 1998b.

C. Furusawa and K. Kaneko. Complex organization in multicellularity as a necessity in evolution. *Artificial Life*, 6:265–281, 2000.

N. Geard and J. Wiles. A gene network model for developing cell lineages. *Artificial Life*, 11:249–267, 2005.

N. L. Geard and J. Wiles. Linmap: Visualizing complexity gradients in evolutionary landscapes. *Artificial Life*, 14:277–297, 2008.

P. Gönczy. Mechanisms of asymmetric cell division: Flies and worms pave the way. *Nature Reviews Molecular Cell Biology*, 9(5):355–366, 2008.

J. Hench, J. Henriksson, M. Lüppert, and T. R. Bürglin. Spatio-temporal reference model of caenorhabditis elegans embryogenesis with cell contact maps. *Developmental Biology*, 333(1):1–13, 2009.

M. Howard-Ashby, S. C. Materna, C. T. Brown, L. Chen, R. A. Cameron, and E. H. Davidson. Gene families encoding transcription factors expressed in early development of Strongylocentrotus purpuratus. *Developmental Biology*, 300:90–107, 2006.

Y. N. Jan and L. Y. Jan. Asymmetric cell division. *Nature*, 392: 775–778, 1998.

T. Kaletta, H. Schnabel, and R. Schnabel. Binary specification of the embryonic lineage in Caenorhabditis elegans. *Nature*, 390: 294–298, 1997.

K. Kaneko. Coupled maps with growth and death: An approach to cell differentiation. *Physica D: Nonlinear Phenomena*, 103: 505–527, 1997.

S. A. Kauffman. Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of Theoretical Biology*, 22:437–467, 1969.

J. F. Knabe, C. L. Nehaniv, and M. J. Schilstra. Genetic regulatory network models of biological clocks: Evolutionary history matters. *Artificial Life*, 14:135–148, 2008.

E. Lécuyer and P. Tomancak. Mapping the gene expression universe. *Current Opinion in Genetics and Development*, 18(6):506–512, 2008.

I. Lee, B. Lehner, C. Crombie, W. Wong, A. G. Fraser, and E. M. Marcotte. A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans. *Nature Genetics*, 40:181–188, 2008.

R. Lohaus, N. Geard, J. Wiles, and R. Azevedo. A generative bias towards average complexity in artificial cell lineages. *Proceedings of the Royal Society B: Biological Sciences*, 274:1741–1750, 2007.

S. C. Materna and E. H. Davidson. Logic of gene regulatory networks. *Current Opinion in Biotechnology*, 18:351–354, 2007.

A. Mochizuki. Structure of regulatory networks and diversity of gene expression patterns. *Journal of Theoretical Biology*, 250:307–321, 2008.

P. Oliveri, Q. Tu, and E. H. Davidson. Global regulatory logic for specification of an embryonic cell lineage. *Proceedings of the National Academy of Sciences of the United States of America*, 105: 5955–5962, 2008.

E. Plahte, T. Mestl, and S. W. Omholt. Global analysis of steady points for systems of differential equations with sigmoid interactions. *Dynamics and Stability of Systems*, 9:275–291, 1994.

U. Platzer and H.-P. Meinzer. Genetic networks in the early development of Caenorhabditis elegans. *International Review of Cytology*, 234:47–100, 2004.

M. J. Schilstra and H. Bolouri. Modelling the regulation of gene expression in genetic regulatory networks, 2003. URL `http://strc.herts.ac.uk/bio/maria/NetBuilder/Theory/NetBuilderTheoryDownload.pdf`.

H. Siebert and A. Bockmayr. Temporal constraints in the logical analysis of regulatory networks. *Theoretical Computer Science*, 391:258–275, 2008.

H. S. Silva and M. L. Martins. A cellular automata model for cell differentiation. *Physica A: Statistical Mechanics and its Applications*, 322:555–566, 2003.

J. Smith, C. Theodoris, and E. H. Davidson. A gene regulatory network subcircuit drives a dynamic pattern of gene expression. *Science*, 318:794–797, 2007.

P. Smolen, D. Baxter, and J. H. Byrne. Modeling transcriptional control in gene networks — Methods, recent results, and future directions. *Bulletin of Mathematical Biology*, 62:247–292, 2000.

J. E. Sulston, E. Schierenberg, J. G. White, and J. N. Thomson. The embryonic cell lineage of the nematode caenorhabditis elegans. *Developmental Biology*, 100:64–119, 1983.

D. J. Taatjes, M. T. Marr, and R. Tjian. Regulatory diversity among metazoan co-activator complexes. *Nature Reviews Molecular Cell Biology*, 5(5):403–410, 2004.

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

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

R. Thomas, D. Thieffry, and M. Kaufman. 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*, 57:247–276, 1995.

A. M. Turing. The chemical basis of morphogenesis. *Philos. Trans. R. Soc. London, Ser. B*, 237:37–72, 1952.

C. Walter, R. Parker, and M. Yčas. A model for binary logic in biochemical systems. *Journal of Theoretical Biology*, 15:208–217, 1967.

Z. Wei, R. C. Angerer, and L. M. Angerer. A database of mRNA expression patterns for the sea urchin embryo. *Developmental Biology*, 300:476–484, 2006.

Wormatlas. `http://www.wormatlas.org`. The entire embryonic *C. elegans* lineage tree can be found at `http://www.wormatlas.org/images/embryoniclineage.jpg`.

H. Yoshida, C. Furusawa, and K. Kaneko. Selection of initial condition for recursive production of multicellular organisms. *Journal of Theoretical Biology*, 233:501–514, 2005.