Automatic Design of Boolean Networks for Modelling Cell Differentiation

Stefano Benedettini[1], Andrea Roli[1], Roberto Serra[2], and Marco Villani[2]

[1] DEIS-Cesena
Alma Mater Studiorum Università di Bologna
[2] Faculty of Mathematical, Physical and Natural Sciences
Università di Modena e Reggio Emilia &
European Centre for Living Technology, Venice

Abstract. A mathematical model based on Random Boolean Networks has been recently proposed to describe the main features of cell differentiation. The model captures in a unique framework all the main phenomena involved in cell differentiation and can be subject to experimental testing. A prominent role in the model is played by cellular noise, which somehow controls the cell ontogenetic process from the stem, totipotent state to the mature, completely differentiated one. Noise is high in stem cells and it decreases while the cell undergoes the differentiation process. A limitation of the current mathematical model is that Random Boolean Networks, as an ensemble, are not endowed with the property of showing a smooth relation between noise level and the differentiation stages of cells. In this work, we show that it is possible to generate an ensemble of Boolean networks that can accomplish such requirement, while keeping the other main relevant statistical features of classical Random Boolean Networks. This ensemble is designed by means of an optimisation process, in which a stochastic local search optimises an objective function which accounts for the requirements the network ensemble has to fulfil.

1 Introduction

Cell differentiation is the process whereby stem cells, which can develop into different types, become more and more specialised. A mathematical model of cell differentiation has been recently proposed by Serra, Villani et al. [1,2]. The model is an abstract one (i.e., it does not refer to a specific organism or cell type) and it aims at reproducing the most relevant features of cell differentiation, which are the following:

(i) there exist different degrees of differentiation, that span from totipotent stem cells to fully differentiated cells;
(ii) there are both *deterministic* differentiation, where signals trigger the progress of multipotent cells into more differentiated types, in well defined lineages, and *stochastic* differentiation, where populations of identical multipotent cells stochastically generate different cell types;

(iii) limited reversibility: differentiation is almost always irreversible, but there are limited exceptions under the action of appropriate signals;
(iv) induced pluripotency: fully differentiated cells can come back to a pluripotent state by modifying the expression of some genes;
(v) induced change of cell type: modification of the expression of few genes can directly convert one differentiated cell type into another.

The differentiation model is based on a noisy version of a well-know model of gene networks, that is, the Random Boolean Network (RBN) model. In spite of the assumption of discreteness, RBNs have been proven to describe important experimental facts concerning gene expression [3,4,5]. The dynamics of "classical" RBNs is discrete and synchronous, so fixed points and cycles are the only possible asymptotic states in finite networks; typically, a single RBN owns more than one attractor. Attractors of RBNs are unstable with respect to noise even at low levels, like, e.g., transient flips of randomly chosen nodes. In fact, even if the flips last for a single time step, one often observes transitions from that attractor to another one. Ribeiro and Kauffman [6] observed that it is possible to identify in the attractors' landscape subsets of attractors, which they called *Ergodic Sets* (ESs), which entrap the system in the long time limit, so the system continues to jump between attractors which belong to the set. Unfortunately, it turns out that most noisy RBNs have just one such set: this observation rules out the possibility to associate them to cell types. The model proposed by Serra, Villani et al. overcomes this problem by observing that flips are a kind of fairly intense noise, as they amount to silencing an expressed gene or to express a gene which would otherwise be inactive: this event may happen with a very low probability in the cell lifetime. It is possible therefore to introduce a threshold θ, and neglect all the transitions whose occurrence probability is lower than θ. In such a way, the notion of ES has to be modified in that of *Threshold Ergodic Set* (TES_θ), a set of attractors linked only by jumps having a probability higher than θ, that entrap the system in the long time limit. A TES_θ is therefore a subset of attractors which are each other directly or indirectly θ-reachable[3] and from which no transition can allow escaping. The threshold clearly is related to the level of noise in the cell, and scales with its reciprocal (the frequency of flips). Hence, cell types are associated to TESs, which represent coherent stable ways of functioning of the same genome even in the presence of noise. According to this framework, RBNs can host more than one TES, avoiding in such a way the problem that hampered the straightforward association of cell types to ES. At high noise level the system can jump among all the attractors, thus modelling stem cells, while as the threshold is increased (i.e., noise is reduced) the cell becomes entrapped in a smaller TES, that represents a multipotent cell. At very high threshold values all the attractors are also TES, a condition likely to describe final cell types. Indeed, there are experimental indications in favour of the key hypothesis that noise in stem cells is higher than in more differentiated ones. In this model, cell differentiation is an emerging property originating from

[3] Reachable by means of transition whose probability exceeds the threshold θ.

the interactions of many genes: its main features therefore should be shared by a variety of different organisms.

This single model is able to capture all the phenomena encompassed by cell differentiation and its application to real cell differentiation processes is open to validations.[4] Nevertheless, the model could be ameliorated in some important aspects. In fact, while for RBNs it is true that the number of TESs increases with the threshold θ, the largest amount of this increase takes place for a very narrow range of values, necessitating in such a way a very sophisticated control to precisely tune the correct threshold for the required differentiation level. This paper presents a way of overcoming this weakness by providing a method for designing BNs such that the increase in TES number happens within a larger threshold range.

This contribution is structured as follows. Section 2 details the limitation of the current model and introduces the revision needed to accomplish for the proper relation between number of TES and threshold. In Section 3, we illustrate the method we used for obtaining such BNs. Section 4 describes the experiments we made and present a statistical analysis of the results. Finally, Section 5 summarises the main outcome of this contribution and outlines future work.

2 Improved model

The mathematical model previously illustrated can capture all the relevant phenomena of cell differentiation. Precise quantitative analyses can be undertaken depending on the availability of experimental data, which unfortunately are scarce and incomplete at the present time. However, the relevance of the model can be assessed in the context of the so-called *ensemble approach* [7], which aims at finding classes of genetic regulatory network models that match statistical features of living cells. In the case of cell differentiation, the model proposed by Serra, Villani et al. succeeds in describing the way in which a lineage tree is hierarchically organised and can also explain the other phenomena involved in the differentiation process from the ensemble approach standpoint. Nevertheless, as already emphasised, it requires a precise control in a very narrow threshold range, resulting in this aspect not completely satisfactory with respect to biological feasibility. In fact, in the ensemble of noisy RBNs considered in the model, the number of TESs varies approximately as a Heaviside step function of θ: one or very few TESs can be found for $\theta \leq \theta_0$ and the maximum number of TESs (equal to the number of attractors) is achieved with θ just above θ_0.[5] This behaviour is prone to errors in identifying the correct differentiation level within the lineage tree, and therefore biologically not plausible. Therefore, we would like to find an ensemble of BNs such that the main properties characterising RBNs are preserved and the number of TESs scales smoothly with the threshold θ. This goal can be achieved by applying a recently proposed method, which

[4] There are some positive but not yet definitive experimental data.

[5] Of course, the property is typical of the ensemble and isolated exceptions could be found. The value θ_0 depends on the specific instance considered.

consists in converting the BN design problem into an optimisation one and solve it through stochastic local search [8]. This automatic design method has been proven to successfully solve BN design problems [9,10,11] and it will be detailed, for the case at hand, in the following section.

3 Methods

The problem of designing a BN or a set of BNs meeting given dynamical requirements can be stated as an optimisation problem. In particular, one has to define the decision variables and the objective function.[6] In principle, the optimisation problem can be solved by any search method; however, stochastic local search have shown to be very effective in tackling these kinds of problems and is thus our preferential choice. For this specific case, we assume that the topology of the network is set initially according to a random model [12] and kept fixed during search. The decision variables of the problem maneuver the Boolean functions of the BN nodes: for BNs with N nodes, each with K inputs, we introduce $2^K N$ Boolean decision variables, which define the transition functions of network nodes. Therefore, the local search can explore the space of all the possible assignments of Boolean functions to the nodes, trying to minimise an objective function which estimates the distance between a current BN instance and the requirements posted.

The local search algorithm employed here is an Iterated Local Search (ILS) embedding a Stochastic Descent (SD). ILS is a well-known stochastic local search framework, successfully applied to many hard combinatorial optimisation problems. In a nutshell, ILS applies a local search to an initial solution until it finds a local optimum; then it perturbs the solution and it restarts local search. An overview on theory and applications of ILS can be found in [8]. ILS makes it possible to combine the efficiency of local search with the capability of escaping from the basin of attraction of local optima. SD is a simple local search in which a neighbour of the current solution is randomly picked and accepted if it is at least as good as the current one. In a sense, our combination of ILS with SD can be seen as an iterated version of an *adaptive walk* in which restart is not random but performed in such a way that diversification is increased gradually. In SD, the move consists of one random flip in the truth table of a node transition function (the node is also chosen randomly). The perturbation used inside ILS is performed as a random flip for every transition function. The same algorithm as been used in a paper by Benedettini et al. [10], in which the authors design an ensemble of BNs with maximally distant attractors. The interested reader is referred to that paper for further details on the algorithm.

3.1 Objective function

The aim of our local search is to find BNs endowed with the two following properties:

[6] We assume that constraints are either implicitly satisfied or they are relaxed and included in the objective function.

1. the number of TESs should grow smoothly with the threshold θ;
2. attractors should be stable, i.e., the probability of transition $a \rightarrow a$, where a is an attractor, should be high. This property ensures that we can put into relation the attractors of the BN with the cell types of completely differentiated cells. Some attractors may be sensitive to small perturbations, but the majority should be stable [13].

The objective function closely reflects the requirements mentioned above. In particular, we opted for a linear relation between the number of TESs and the threshold θ, which is the simplest choice, yet effective. The computation of the objective function requires first the calculation of the *transition graph*, i.e., a directed graph whose vertices are attractors and edges represent transitions between attractors. Edges are weighted with transition probabilities. The transition graph $\mathcal{G}$ is calculated with the algorithm specified in [2]. The objective function consists of the following two terms:

Attractor stability: the first contribution to the objective function is given by a term S calculated as the fraction of vertices in $\mathcal{G}$ with a self-loop with weight greater then or equal to 0.8.

Number of TESs as a linear function of θ: the second contribute E (E stands for error as we see in Equation 1) is calculated as follows: let us select a sequence Θ of n equally spaced values from interval $[0, 0.5[$, i.e., $\Theta = 0, \frac{1}{2n}, \frac{2}{2n}, \ldots, \frac{n-1}{2n}$. Let us also define the sequence $s_i, 0 \leq i \leq n$ as the number of TES_{Θ_i} (TESs with threshold Θ_i) in $\mathcal{G}$. Contribute E is:

$$\sum_{i=0}^{n} |\frac{s_i}{s_n} - \frac{i}{n}|. \tag{1}$$

Objective function: the objective function to be *minimised* is

$$(2 - S)E \tag{2}$$

Let us motivate our choices. Contribute S directly reflects the requirement on attractor stability.We should make clear that resulting networks do not necessarily have transition graphs with self-loops of weight 0.8, but they are forced to have most of the attractors with this property. Contribute E addresses the first requirement; basically, we ask for a sequence s_i that is as *smooth* as possible, i.e, we want that s_i gradually grows to its maximum value s_n, the linear growth we are using in this paper being the simplest option among a more ample set of possibilities. In Equation 1 we divide s_i by s_n so that E is not dependent on the number of TESs. Finally, the two contributes are composed so that S takes the role of a *penalty*: the smaller S, the greater the increase on the error E.

We wish to conclude this section with some remarks on our design choices. For what the objective function is concerned, we would like to observe that, although the proposed linear model of growth of the number of TESs is not based on any particular biological hypothesis or experimental observations, our objective function is an effective criterion that guides our local search algorithm

towards promising area of the solution space. Although this claim should be supported by a more detailed analysis of the search space (left for future work), an empirical proof of its effectiveness is given in the next section. Moreover, the choice of some parameters, like for example the values 0.8 (the self-loop desired weight) or 0.5 (the interval length spanned by sequence Θ), is arbitrary; evaluating the robustness of our results with respect to variations in these parameters lies beyond the purpose of the present work and will be the aim of further investigations.

Finally, an algorithmic detail. From a graph theoretical point of view, the number of TESs can be calculated as follows: first we remove from $\mathcal{G}$ all edges with weight less than θ, than we compute the *condensation* of $\mathcal{G}$ [14] and we count the vertices with null out-degree.

4 Results

Typical RBNs are characterised by constant input connectivity K and Boolean functions chosen at random with on average $2^K p$ true entries in the truth table, where $p \in [0, 1]$ is called *bias*. Depending on the values of K and p the dynamics of RBNs is *ordered* or *disordered* (also called *chaotic*, with a slight abuse of terms). In the first case, the majority of nodes in the attractors is frozen; any moderate-size perturbation is rapidly dampened and the network returns to its original attractor. Conversely, in disordered dynamics, attractor cycles are very long and the system is extremely sensitive to small perturbations: slightly different initial states lead to exponentially diverging trajectories in the state space. RBNs temporal evolution undergoes a second order phase transition between order and chaos, governed by the following relation between K and p: $K_c = [2p_c(1-p_c)]^{-1}$, where the subscript c denotes the critical values [15]. Networks along the *critical* line are the ones with the maximal match with living cell features [16,5].

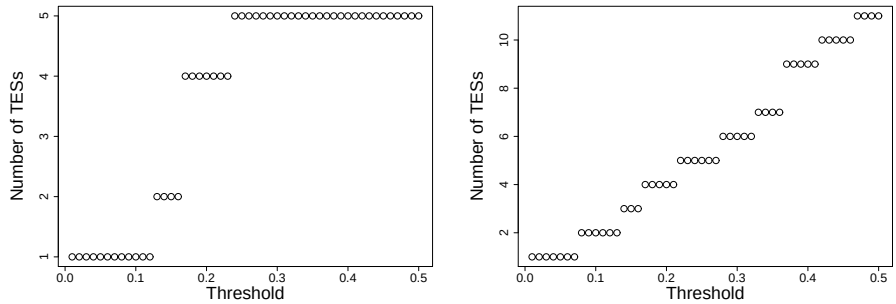
We tested our algorithm on two test sets, both composed of *critical* RBNs with $N = 100$ nodes and constant in-degree. The first test set consists of 30 critical RBNs with in-degree $K = 2$ (whence $p = 0.5$); the second test set contains 30 critical RBNs with in-degree $K = 3$ (whence $p \approx 0.788$). Networks in these two ensembles constitute the initial solutions of our local search and will be collectively referred to as *initial ensemble*. Similarly, the set of BNs obtained after running our local search constitutes our *optimised ensemble*.

In order to compute our objective function we had to compute the transition graph. We initialise the algorithm with attractors found after a sample of 1000 initial conditions (more attractors may of course be found during algorithm execution and are recursively considered in the algorithm). We considered only trajectories with at most 1500 steps: if an attractor is not found in this number of steps, the sample is discarded. As for ILS is concerned, we set a runtime limit of 3 hours per experiment.

4.1 Analysis of network properties

In order to analyse a BN we sampled its state space in 100000 random initial conditions, since an exhaustive test would be prohibitive. For each network, we recorded the number of attractors, their relative basin sizes and their periods. In addition, we computed the transition graph and the number of TESs sequence, as previously explained.

The first remarkable result is that the networks designed through our local search have the number of TESs which smoothly increases with the threshold. A typical case is depicted in Figure 1, where we can observe that the number of TESs increases within a wide threshold range. The transition graph corresponding to this automatically designed network is drawn in Figure 2. This property is common to almost all the networks generated by the search procedure and it can be considered as an invariant of the ensemble.[7]



(a) Typical random BN (b) Automatically designed BN

Fig. 1: Number of TES as a function of the threshold in a typical random BN (Figure 1a) and an automatic designed one (Figure 1b).

Statistics on attractors stability, weights in the transition graph, number of attractors, their relative basin sizes and their periods are summarised by boxplots in Figures 3, 4 and 5. Each figure depicts the distribution of a measure on the initial ensemble (left boxplot) and the optimised ensemble (right boxplot) for all test sets.

Figure 3 shows two measures that try to characterise weights of the transition graph and thereby statistically demonstrate the effectiveness of our method in designing BNs with the characteristics stated in Section 3.1. The transition graph can be represented by a weight matrix $(p^i_j), 0 \leq p^i_j \leq 1$, where p^i_j represents the probability of the network to go from attractor i to the basin of attractor j after

[7] Since the search process is stochastic, we can consider our design method as a biased sampling in the space of BNs.

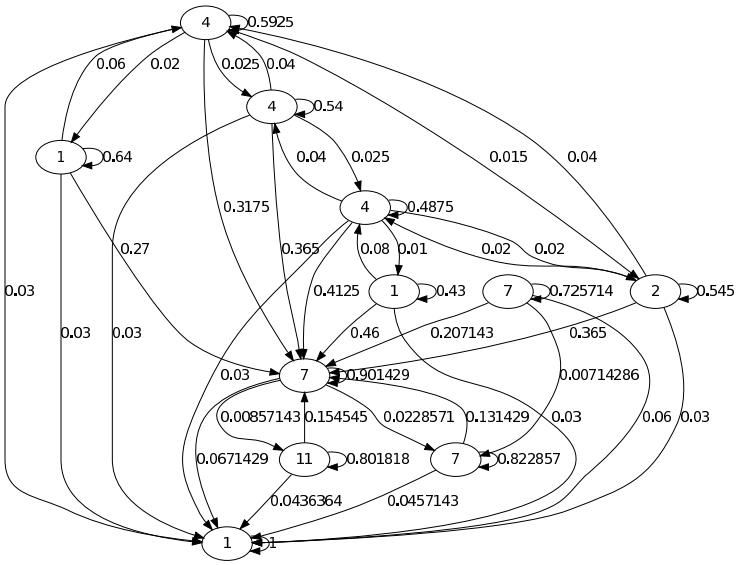
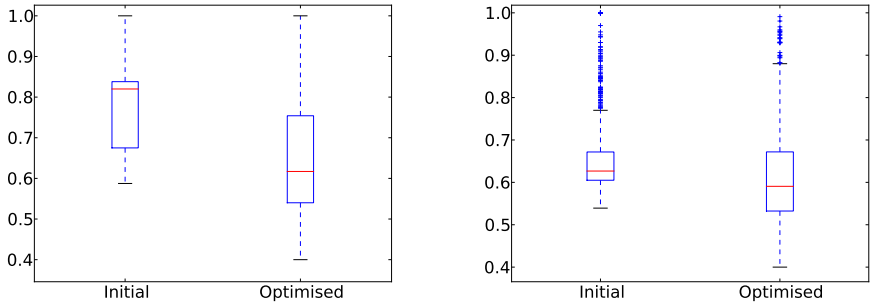


Fig. 2: Attractor transition graph for an automatic designed BN.

a random flip. Self-loops weights p_i^i indicate how insensitive is attractor i to random flips; in general we want this probability to be high. Attractor stability is a non-scalar measure of a BN, that is, for a single network we have more than one value (actually one for each attractor).

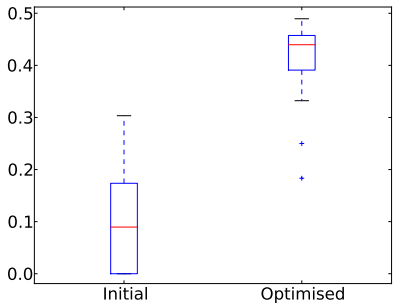
Figures 3a and 3b describe attractor stability distribution of networks in the initial and optimised ensembles; since this measure is not scalar, to draw our boxplots we merged together the data for all networks in each ensemble. Figures 3c and 3d characterise instead the distribution of the non self-loop weights (once again a non-scalar measure). For each BN, we collect in a vector P all non-zero elements of (p_j^i) outside the main diagonal; afterwards, we compute the difference $\max P - \min P$. Figures 3c and 3d show the distribution of such difference for initial and optimised networks. It can be seen that attractor stability is lower but close to the initial ensemble (requirement *(1)* in Section 3.1); at the same time, edge weights $p_j^i, i \neq j$ are more spread out, in accordance with our objective (requirement *(2)* in Section 3.1).

Figure 4 depicts the distribution of number of attractors. As a result of the search process, the number of attractor does not vary in a statistically significant way, although it seems that it tends to grow a bit. Figure 5 shows basin sizes and attractor periods. Since these two measures are not scalar, boxplots in Figure 5 summarise the distribution of *median* basin size and attractor period calculated on each BN. We see that attractor period does not statistically vary, but the distribution of basin sizes is remarkably different; specifically, the search process tends to shrink basin sizes. Intuitively, we can say that attractors with small basin sizes are likely to be less robust than attractors with larger basin sizes

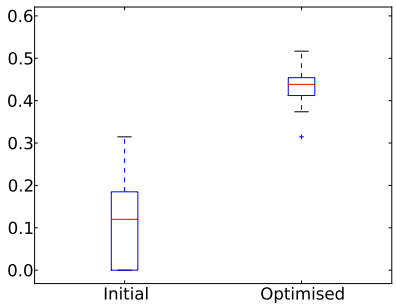


(a) Attractor stability ($K = 2$)

(b) Attractor stability ($K = 3$)



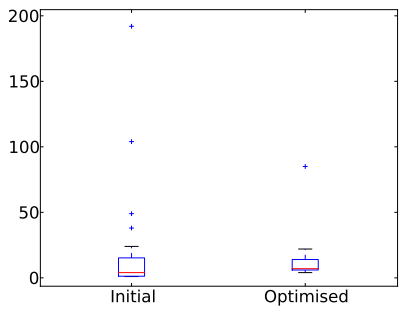
(c) Transitions ($K = 2$)



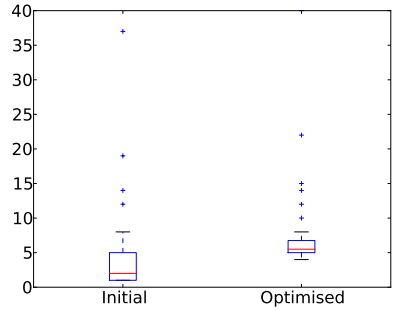
(d) Transitions ($K = 3$)

Fig. 3: Attractor stability and transitions for initial and optimised BNs in all test sets. Figures 3a and 3b depict the distribution of attractor stability across all networks in either test set; the distribution is constructed by merging all data samples gathered. Figure 3c and 3d characterise the distribution of edge weights in the transition graph as explained in text.

simply because the basins of the latter contain more states. Therefore, one would expect that networks with smaller basins (feature typical of optimised networks) are also characterised by unstable attractors. However, Figures 3a and 3b clearly show that attractor stability is essentially unchanged. From these data we can conclude that the search process reorganises the basins of the attractors in such a way as to satisfy our stated requirements: the basins are therefore "rebalanced" so as to have generally stable attractors. It appears that, in order to achieve this goal, our local search had to reduce the size of some of the larger basins. This is the motivation why the median stability in Figures 3a and 3b slightly decreases (about a 0.2 decrease for networks with $K = 2$ and a decrement less than 0.1 for networks with $K = 3$).

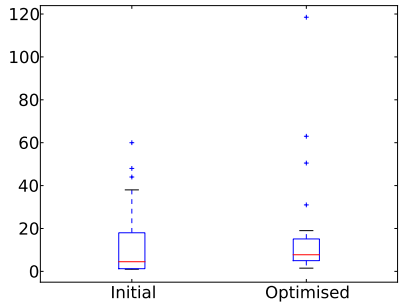


(a) Number of attractors ($K = 2$)

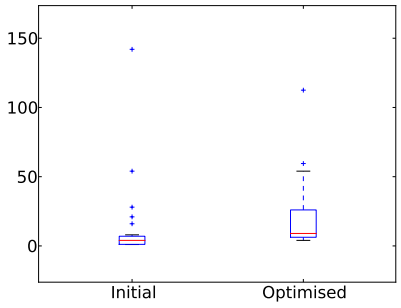


(b) Number of attractors ($K = 3$)

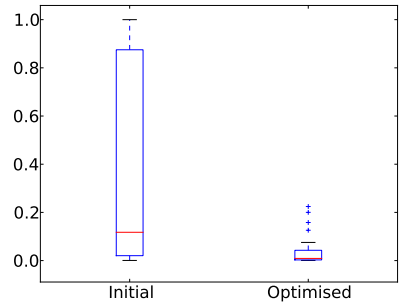
Fig. 4: Distributions of the number of attractors (Figures 4a and 4b) on initial and optimised networks.



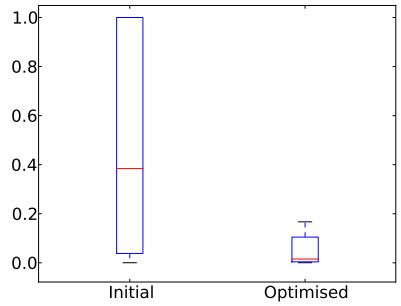
(a) Attractor periods ($K = 2$)



(b) Attractor periods ($K = 3$)



(c) Attractor basins ($K = 2$)



(d) Attractor basins ($K = 3$)

Fig. 5: Figures 5a through 5d summarise the distribution of *median* attractor period (5a and 5b) and *median* normalised basin size (5c and 5d) respectively.

5 Conclusion and future work

In this paper we have proposed an improvement of a mathematical model for cell differentiation that makes use of RBNs. An ensemble of BNs that match the dynamical requirements deriving from biological plausibility has been designed by means of an optimisation process that make use of stochastic local search. The BNs generated are characterised by a more realistic relation between the number of TESs and the threshold, conserving the other relevant properties of the RBN ensemble. In particular, results show that in the ensemble generated by the optimisation process the number of TESs grows smoothly with the threshold and attractors are robust.

Future work will address the extension of the model by introducing in the optimisation process further features of cell differentiation, such as properties concerning deterministic and stochastic differentiation. Furthermore, besides the ensemble approach, we are planning to validate the model against experimental data collected for specific organisms.

References

1. R. Serra, M. Villani, A. Barbieri, S.A. Kauffman, and A. Colacci. On the dynamics of random Boolean networks subject to noise: Attractors, ergodic sets and cell types. *Journal of Theoretical Biology*, 265(2):185–193, 2010.
2. M. Villani, A. Barbieri, and R. Serra. A dynamical model of genetic networks for cell differentiation. *PLoS ONE*, 6(3):e17703:1–9, 03 2011.
3. R. Serra, M. Villani, and A. Semeria. Genetic network models and statistical properties of gene expression data in knock-out experiments. *Journal of Theoretical Biology*, 227:149–157, 2004.
4. M. Aldana, E. Balleza, S.A. Kauffman, and O. Resendiz. Robustness and evolvability in genetic regulatory networks. *Journal of Theoretical Biology*, 245:433–448, 2007.
5. E. Balleza, E.R. Alvarez-Buylla, A. Chaos, S.A. Kauffman, I. Shmulevich, and M. Aldana. Critical dynamics in genetic regulatory networks: Examples from four kingdoms. *PLoS ONE*, 3(6):e2456, 06 2008.
6. A.S. Ribeiro and S.A. Kauffman. Noisy attractors and ergodic sets in models of gene regulatory networks. *Journal of Theoretical Biology*, (247):743–755, 2007.
7. S.A. Kauffman. A proposal for using the ensemble approach to understand genetic regulatory networks. *Journal of Theoretical Biology*, 230:581–590, 2004.
8. H.H. Hoos and T. Stützle. *Stochastic Local Search: Foundations and Applications*. Morgan Kaufmann Publishers, San Francisco, CA, 2005.
9. A Roli, C. Arcaroli, M. Lazzarini, and S. Benedettini. Boolean networks design by genetic algorithms. In M. Villani and S. Cagnoni, editors, *Proceedings of CEEI 2009 - Workshop on complexity, evolution and emergent intelligence*, Reggio Emilia, Italy, 2009.
10. S. Benedettini, A. Roli, R. Serra, and M. Villani. Stochastic local search to automatically design Boolean networks with maximally distant attractors. In C. Di Chio, S. Cagnoni, C. Cotta, M. Ebner, A. Ekárt, A. Esparcia-Alcázar, J. Merelo, F. Neri, M. Preuss, H. Richter, J. Togelius, and G. Yannakakis, editors, *Applications of Evolutionary Computation*, volume 6624 of *Lecture Notes in Computer Science*, pages 22–31. Springer, Heidelberg, Germany, 2011.

11. A. Roli, M. Manfroni, C. Pinciroli, and M. Birattari. On the design of Boolean network robots. In C. Di Chio, S. Cagnoni, C. Cotta, M. Ebner, A. Ekárt, A. Esparcia-Alcázar, J. Merelo, F. Neri, M. Preuss, H. Richter, J. Togelius, and G. Yannakakis, editors, *Applications of Evolutionary Computation*, volume 6624 of *Lecture Notes in Computer Science*, pages 43–52. Springer, Heidelberg, Germany, 2011.
12. S.A. Kauffman. *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford University Press, UK, 1993.
13. F. Li, T. Long, Y. Lu, Q. Ouyang, and C. Tang. The yeast cell-cycle network is robustly designed. *PNAS*, 101(14):4781–4786, 2004.
14. Wikipedia. Strongly connected component — Wikipedia, The Free Encyclopedia, 2011. [Online; accessed February 7, 2012].
15. B. Derrida and Y. Pomeau. Random networks of automata: a simple annealed approximation. *Europhysics Letters*, 1(2):45–49, 1986.
16. I. Shmulevich, S.A. Kauffman, and M. Aldana. Eukaryotic cells are dynamically ordered or critical but not chaotic. *Proceedings of the National Academy of Sciences of the United States of America*, 102(38):13439–13444, 2005.