

# A SAT-Based Algorithm for Computing Attractors in Synchronous Boolean Networks

Elena Dubrova and Maxim Teslenko

Royal Institute of Technology (KTH), Electrum 229, 164 46 Kista, Sweden  
{dubrova,maximt}@kth.se

**Abstract.** This paper addresses the problem of finding cycles in the state transition graphs of synchronous Boolean networks. Synchronous Boolean networks are a class of deterministic finite state machines which are used for the modeling of gene regulatory networks. Their state transition graph cycles, called *attractors*, represent cell types of the organism being modeled. When the effect of a disease or a mutation on an organism is studied, attractors have to be re-computed every time a fault is injected in the model. We present an algorithm for finding attractors which uses a SAT-based bounded model checking. Novel features of the algorithm compared to the traditional SAT-based bounded model checking approaches are: (1) a termination condition which does not require an explicit computation of the diameter and (2) a technique to reduce the number of additional clauses which are needed to make paths loop-free. The presented algorithm uses much less space than existing BDD-based approaches and has a potential to handle several orders of magnitude larger networks.

**Keywords:** bounded model checking, SAT, Boolean network, attractor, gene regulatory network

## 1 Introduction

A gene regulatory network (GRN) is a collection of DNA segments in a cell, called *genes*, which interact with each other [1]. Each gene contains information that determine what the gene does and when the gene is active, or *expressed*. When a gene is active a process called *transcription* takes place, producing an *ribonucleic acid (RNA)* copy of the gene's information. This piece of RNA can then direct the synthesis of *proteins*. RNA or protein molecules resulting from the transcription process are known as *gene products*.

Mathematical models of GRNs have been developed to capture the behavior of organisms being modeled. Common GRN models include ordinary and partial differential equations, Boolean networks and their generalizations, Petri nets, Bayesian networks, stochastic equations, and process calculi [2]. There is always tension between generality and level of details, and thus tractability, of a model. The most appropriate mathematical framework can be selected depending on the scale involved, the nature of the available information, and the problem studied. In this paper, we consider the Boolean

network model, which has been shown useful for exploring GRNs in the context of cellular differentiation, cell cycle regulation, immune response, and evolution (see [3] for an overview).

The *Boolean network* is a discrete-space discrete-time model in which every gene is viewed as a vertex whose input values represent gene products and output values represent the level of gene expression [4]. The edges between vertices represent the interactions between genes. These interactions can be *activatory*, with an increase in the concentration of gene products in one gene leading to an increase in the level of gene expression in other gene, or *inhibitory*, with an increase in one leading to a decrease in the other. The nature of influence of regulators on a gene is reflected by the Boolean function assigned to the vertex. In this paper we consider the synchronous type of Boolean networks in which the values of functions of all vertices are updated simultaneously at each time step.

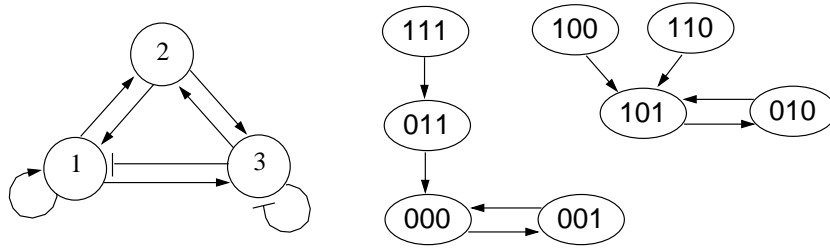
Synchronous Boolean networks can be considered as a class of deterministic finite state machines. Any sequence of consecutive states of a network eventually converges to either a single state, or a cycle of states, called *attractor*. Attractors represent the pattern of gene expressions in the corresponding cell types of the organism being modeled [5]. When the effect of a disease or a mutation on an organism is studied, attractors have to be re-computed every time a fault is injected in the model [3].

All algorithms for computing attractors in Boolean networks face a state-space explosion problem that must be addressed to handle large-scale models. A common approach to combat it is to use symbolic algorithms which avoid building the state transition graph describing the dynamic behavior of a GRN. Instead, the state transition graph is represented implicitly by means of Binary Decision Diagrams (BDDs) [6]. Algorithms based on BDDs [7–9] are usually able to process GRN models with up to a hundred of state variables. However, for larger networks, BDDs become too memory-consuming. Simulation-based approaches [10–12] can be applied to large networks, however, they are incomplete.

Propositional decision procedures (SAT) do not suffer from the potential space explosion of BDDs and can handle propositional satisfiability problems with thousands of variables [13]. This work is the first step in applying SAT procedures to computing attractors. The presented approach is based on SAT-based bounded model checking [14]. We use a SAT-solver for identification of paths of a particular length  $k$  in the state transition graph of a Boolean network. First we generate a propositional formula representing an unfolding of the transition relation of the network for  $k$  steps. A satisfying assignment to this propositional formula corresponds to a valid path in the state transition graph. The process is repeated iteratively for larger and larger values of  $k$  until all attractors are identified.

Note that systems which we are dealing with are deterministic and therefore the problem we are addressing is simpler than the traditional bounded model checking. This allows us to use a simple termination condition which does not require an explicit computation of the diameter and also to reduce the number of additional clauses which are needed to make paths loop-free.

This paper contributes to the ongoing work on finding attractors by providing a complete solution which uses much less space than BDD-based approaches and thus



**Fig. 1.** A 3-vertex Boolean network (left) and its state transition graph (right).

has a potential to handle several orders of magnitude larger networks. The existing simulation- and BDD-based algorithms for finding attractors are only applicable to simple organisms such as the yeast, a flower, or a fruit fly. The presented approach opens a possibility for exploring much more complex organisms including humans.

The paper is organized as follows. Section 2 gives a background on synchronous Boolean networks. In Section 3 we describe the intuitive idea behind the presented approach. Section 4 presents the algorithm. Section 5 summarizes experimental results. Section 6 concludes the paper and discusses open problems.

## 2 The Boolean Network Model

In the *Boolean network* model of GRNs [4], every gene is represented by a vertex  $i$  in a directed graph with an associated state variable  $x_i$  that takes the value 1 if the gene is expressed and 0 otherwise. An edge from one vertex to another indicates that the former gene regulates the latter. Each vertex  $i$  also has an associated Boolean function  $f_i$  which reflects the nature of influence of its regulators.

Time is viewed as proceeding in discrete steps. For the *synchronous* type of update, the expressions of all genes are changed simultaneously. At each time step, the next value of the variable  $x_i$  is determined by current values of  $r$  regulators of  $i$  as:

$$x_i^+ = f_i(x_{i_1}, x_{i_2}, \dots, x_{i_{r_i}}). \quad (1)$$

where  $x_{i_1}, x_{i_2}, \dots, x_{i_{r_i}}$  are state variables associated to the regulators of  $x_i$ .

The *state of the network* is defined as an ordered  $n$ -tuple of state variables  $x_1, x_2, \dots, x_n$  at a particular moment of time. Since a synchronous Boolean network is deterministic and finite, any sequence of its consecutive states eventually converges to either a single state, or a cycle of states, called *attractor*.

An example of a Boolean network with 3 vertices is shown on the left-hand side of Figure 1. Arrows indicate activatory regulation and blunt-ends represent inhibitory

regulation<sup>1</sup> The following Boolean functions are associated to vertices:

$$\begin{aligned} f_1 &= \neg x_3 \wedge (x_1 \vee x_2) \\ f_2 &= x_1 \wedge x_3 \\ f_3 &= \neg x_3 \vee (x_1 \wedge x_2) \end{aligned}$$

where “ $\wedge$ ”, “ $\vee$ ”, and “ $\neg$ ” stand for the Boolean AND, OR and NOT operations, respectively. The State Transition Graph (STG) of this network is shown in the right-hand side of Figure 1. It has two attractors of length two.

The Boolean network model has been extended to a multiple-valued one, in which each variable can take  $m$  values rather than only two [15, 16]. Since each  $m$ -valued variable can be encoded by  $\lceil \log_2 m \rceil$  Boolean variables, any multiple-valued network can be translated to a Boolean one and treated using the presented approach with no conceptual difference.

### 3 Intuitive Idea

The presented algorithm searches for a path of a given length  $k$  in the STG of a Boolean network. If a path is found, we check whether it contains a loop or not. Since each state of the STG of a Boolean network has a unique next state, once a path reaches a loop, it never leaves it. Therefore, we can determine the presence of a loop simply by checking whether the last state of the path occurs at least twice.

Clearly, all states between any two occurrences of the last state belong to a loop. A loop corresponds to an attractor. We mark all attractor’s states. In the following iterations, we will only search for paths in which the last state is not marked.

Until at least one attractor remains unmarked, we can find a path of any length since we can cycle in an attractor forever. However, once all attractors are identified and marked, we will only be able find paths which are shorter than a given length (at most the diameter of the STG). So, when we search for a path of some length  $k$  and it does not exist, this means that all attractors are already identified and the algorithm can terminate.

If a path of length  $k$  does exist and it is loop-free, we double  $k$  and continue the search for a path of the new length.

We illustrate the algorithm on the example of the Boolean network in Figure 1. The algorithm starts from searching for a path of length  $k = 3$ . Suppose that the path we found is  $111 \rightarrow 011 \rightarrow 000$ . Since the last state (000) occurs only once, this path is loop-free. We increase  $k$  to 6 and continue the search for a path of length 6. Suppose that the path we found is  $110 \rightarrow 101 \rightarrow 010 \rightarrow 101 \rightarrow 010 \rightarrow 101$ . Now, we can see that (101, 010) is a two-state attractor and mark it. The following search for a path of length 6 may return us the path  $011 \rightarrow 000 \rightarrow 001 \rightarrow 000 \rightarrow 001 \rightarrow 000$ . Again, we mark (000, 001) is a two-state attractor. The next search shows that there exist no more paths of length 6. We conclude that all attractors are identified and terminate the algorithm.

As another possibility, while searching for a path of length  $k = 3$  we may find a path  $001 \rightarrow 000 \rightarrow 001$ . In this case, we mark (000, 001) is a two-state attractor and continue

<sup>1</sup> The use of two types of edges is common for describing Boolean networks in spite of the fact that it is redundant since the type of regulation is fully defined by the associated functions.

the search for a path of length 3. Next we may find the path  $010 \rightarrow 101 \rightarrow 010$ , and mark it. The following search will show that there exist no more paths of length 3 and algorithm will terminate.

As we can see, the presented algorithm may terminate either before or after the depth of unfolding becomes equal to the diameter of the STG.

## 4 Description of the Algorithm

The pseudocode of the presented algorithm is given as Algorithm 1. We use a SAT-solver for identification of paths of a particular length  $k$  in the STG of a Boolean network. First we generate a propositional formula  $F$  representing an unfolding of the transition relation  $T$  of the network for  $k$  steps. A satisfying assignment to this propositional formula corresponds to a valid path in the STG. The process is repeated iteratively for larger and larger values of  $k$  until all attractors are identified.

### 4.1 Initial unfolding

Given a Boolean network with  $n$  vertices and the transition relation  $T$ , the algorithm first unfolds the transition relation  $k$  times, where  $k = \min(n, 100)$ . We empirically found it more time-efficient to unfold the transition relation directly by  $n$  steps for small networks of size  $n < 100$ . For large networks with  $n > 1000$ , unfolding by  $n$  steps might take too much memory and it is usually unnecessary for identification of all attractors. This is justified by some specific features of gene regulatory networks which we describe in Section 5.

### 4.2 Direction of unfolding

In the pseudocode, we use  $T_{p..r}$  to denote the transition relation  $T$  which is unfolded from the time step  $p$  to the time step  $r$ , i.e.

$$T_{p..r} = \bigwedge_{i=p}^{r-1} T(s_i, s_{i+1}).$$

where  $s_i$  denotes the state of a Boolean network at the time step  $i$ .

One specific feature of our algorithm is that we always unfold  $T$  from some time step  $-p$  to the time step 0 so that the previous time frames rather than the next ones are added to the unfolding. The depth of the unfolding is increased by decreasing  $-p$ . In this way, the last state of the unfolded transition relation is always  $s_0$ , independently of the depth of the unfolding. Later we will explain how this helps us to reduce the number of additional clauses which are needed to make paths loop-free.

### 4.3 Identification of paths

Once the transition relation is unfolded, a SAT-solver is called to find a satisfying assignment for the resulting propositional formula  $F$ . The function Sat in the pseudocode

---

**Algorithm 1** An algorithm for computing attractors in a Boolean network with  $n$  vertices and the transition relation  $T$ .

---

```

i = n
attractor_is_found = False
A(s0) = 0 /* A(s0) is the set of states of all attractors expressed in terms of variables of s0 */
F = T-i..0 /* F is the propositional formula representing the unfolding T-i..0 */
while Sat(F) do
  for (j = -1; j ≥ -i; j --) do
    if sj = s0 then
      for (k = 0; k > j; k --) do
        A(s0) = A(s0) ∪ (s0 ↔ ck) /* ck ∈ {0, 1}n is an assignment of the variables */
                               /* of sk returned by the SAT-solver; s0 ↔ ck is defined by (2) */
      end for
      F = F ∧ ¬A(s0)
      break
    end if
  end for
  if attractor_is_found then
    attractor_is_found = False
  else
    F = F ∧ T-2*i..-i
    i = 2 * i
  end if
end while

```

---

corresponds to a call to a SAT-solver. Sat takes an expression and returns True if there exists an assignment of variables which make the whole expression true.

If a satisfying assignment does not exist, this means that there is no path of length  $i$  in the STG. This implies that all attractors have been already identified and marked in the STG, so the algorithm terminates.

#### 4.4 Checking paths for loops

If a SAT-solver finds a satisfying assignment, the algorithm checks whether there is a loop in the path corresponding to this assignment. As we mentioned before, we can determine the presence of a loop by checking whether the last state of a path occurs at least twice. Since in our case the last state of any unfolded transition relation is  $s_0$ , to identify an attractor, it is sufficient to check whether  $s_0$  occurs at least twice on a path.

#### 4.5 Adding restrictions to $F$

If  $s_j = s_0$  for some  $j \in \{-i, \dots, -2, -1\}$ , then we can conclude that we found an attractor of length  $j$ . In this case, each of the attractor's states is added to the characteristic function  $A(s_0)$  which represents the set of states of all attractors expressed in terms of variables of  $s_0$ .

Benchmark name	vertices $n$	Attractors number $\times$ length	BDD-based [7]		SAT-based		
			sec	MB	sec	MB	unfolding depth
Arabidopsis thaliana	15	$10 \times 1$	0.077	19.14	0.035	1.76	15
Budding yeast	12	$7 \times 1$	0.109	19.82	0.046	1.91	24
Drosophila melanogaster	52	$7 \times 1$	-	> 1000	0.093	2.32	52
Fission yeast	10	$13 \times 1$	0.062	19.04	0.030	1.78	10
Mammalian cell	10	$1 \times 1, 1 \times 7$	0.060	19.04	0.028	1.76	10
T-cell receptor	40	$8 \times 1, 1 \times 6$	0.093	19.34	0.030	1.98	40
T-helper cell	23	$3 \times 1$	0.107	19.61	0.042	1.81	23

**Table 1.** Experimental results for the Boolean networks models of real organisms; ”-” stands for memory blow up.

The  $n$ -bit vector  $c_i \in \{0, 1\}^n$  is used to denote an assignment of variables of  $s_k$  which is returned by the SAT-solver. The notation  $s_0 \leftrightarrow c_k$  means that

$$s_0 \leftrightarrow c_k = \bigwedge_{i=0}^{n-1} (s_0[i] \leftrightarrow c_k[i]), \quad (2)$$

where  $s_0[i]$  is  $i$ th variable of the state  $s_0$  and  $c_k[i]$  is  $i$ th bit-position of the vector  $c_i$ .

By adding  $\neg A(s_0)$  to the propositional formula  $F$  we constrain  $F$  in such a way that any satisfying assignment for  $F$  will contain no states of already identified attractors. Note that, in our case, it is sufficient to ensure that the state  $s_0$  does not belong to any already identified attractor in order to guaranty that no state in this path belongs to an identified attractor. Now it becomes evident why we have chosen to make the last state of any unfolded transition relation  $s_0$ .

## 5 Experimental Results

We have implemented an experimental tool based on the presented algorithm. In this section, we compare it to the BDD-based tool for finding attractors<sup>2</sup> from [7]. Our implementation uses MiniSAT SAT-solver [17]. All experiments were run on a PC with Pentium III 750 MHz processor and 256 Mb memory.

As benchmarks<sup>3</sup> we use existing Boolean networks models of real organisms shown in Table 1: control of flower morphogenesis in *Arabidopsis thaliana* [18], budding yeast cell cycle regulation [19], *Drosophila melanogaster* segment polarity genes expression patterns prediction [20], fission yeast cell cycle regulation [21], the mammalian cell cycle regulation [22], T-cell receptor signaling pathway analysis [23], and T-helper cell differentiation [24].

As we can see from Table 1, the presented SAT-based algorithm uses an order of magnitude less space than the BDD-based algorithm. For the *Drosophila melanogaster*,

<sup>2</sup> Both tools are available at <http://web.it.kth.se/~dubrova>.

<sup>3</sup> At present there is no common set of benchmarks for the GRN simulation tools. We have constructed the input descriptions of models shown in Table 1 manually from the data in the corresponding papers.

the BDD-based algorithm runs out of memory. The memory blow up occurs while trying to construct the initial transition relation  $T$ .

We can also see that for only one benchmark, budding yeast, the depth of unfolding had to be doubled to  $2n$ . For the rest of benchmarks, all attractors were identified after the first unfolding by  $n$  steps.

The performance of the presented algorithm is determined by the number and length of attractors in a network, as well as to the length of the longest path to an attractor. For large networks, we may expect the number of attractors to be considerably smaller than the number of vertices  $n$  of the network. This is because the number of vertices  $n$  in a Boolean network corresponds to the number of relevant genes in the organism it models, and the number of attractors  $N_a$  corresponds to the number of cell types of this organisms. Different hypotheses have been made suggesting that, for large Boolean networks,  $N_a = O(\sqrt{n})$  [5] or  $N_a = O(n^{2/3})$  [11].

As we can see from Table 1, the length of attractors is usually one. This is because the states of attractors represent the expression levels of genes in a given cell type, which are normally stable [20]. So, we may expect the length of attractors to be a small constant for large networks as well.

Finally, the longest path to an attractor is related to the time which takes a cell to settle down into a stable pattern in the process of cell differentiation [5]. For all benchmarks in Table 1 this parameter was smaller than  $n$ . No empirical results are known for larger networks.

## 6 Conclusion

This work is the first step in applying SAT procedures to finding attractors in Boolean networks. We believe that the presented approach has a potential to handle several orders of magnitude larger networks than the ones which can be handled by the BDD-based approaches. Unfortunately, existing Boolean models of real organisms are small, so they do not allow us to support this claim. Our next step is to work in collaboration with biologists on applying the presented tool to create larger models of more complex organisms.

## References

1. B. Alberts, D. Bray, J. Lewis, M. Ra, K. Roberts, and J. D. Watson, *Molecular Biology of the Cell*. New York: Garland Publishing, 1994.
2. T. Schlitt and A. Brazma, "Current approaches to gene regulatory network modelling," *BMC Bioinformatics*, vol. 8, no. Suppl 6, p. S9, 2007.
3. M. Aldana, S. Coopersmith, and L. P. Kadanoff, "Boolean dynamics with random couplings." <http://arXiv.org/abs/adap-org/9305001>.
4. S. A. Kauffman, "Metabolic stability and epigenesis in randomly constructed nets," *Journal of Theoretical Biology*, vol. 22, pp. 437–467, 1969.
5. S. A. Kauffman, *The Origins of Order: Self-Organization and Selection of Evolution*. Oxford: Oxford University Press, 1993.



6. R. Bryant, "Graph-based algorithms for Boolean function manipulation," *Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 35, pp. 677–691, August 1986.
7. E. Dubrova, M. Teslenko, and A. Martinelli, "Kauffman networks: Analysis and applications," in *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 479–484, November 2005.
8. A. Garg, I. Xenarios, L. Mendoza, and G. De Micheli, "An Efficient Method for Dynamic Analysis of Gene Regulatory Networks and in silico Gene Perturbation Experiments," in *11th Annual International Conference on Research in Computational Molecular Biology, RECOMB 2007*, Lecture Notes in Computer Science, Springer, 2007.
9. A. Naldi, D. Thieffry, and C. Chaouiya, "Decision diagrams for the representation and analysis of logical models of genetic networks," in *Computational Methods in Systems Biology*, vol. 4695 of *Lecture Notes in Computer Science*, pp. 233–247, Springer, 2007.
10. A. Wuensche, "The DDLab manual," 2000. [http://www.cogs.susx.ac.uk/users/andywu/man\\_contents.html](http://www.cogs.susx.ac.uk/users/andywu/man_contents.html).
11. S. Bilke and F. Sjunnesson, "Stability of the Kauffman model," *Physical Review E*, vol. 65, p. 016129, 2001.
12. J. E. S. Socolar and S. A. Kauffman, "Scaling in ordered and critical random Boolean networks." <http://arXiv.org/abs/cond-mat/0212306>.
13. M. Davis and H. Putnam, "A computing procedure for quantification theory," *J. ACM*, vol. 7, no. 3, pp. 201–215, 1960.
14. A. Biere, A. Cimatti, E. Clarke, M. Fujita, and Y. Zhu, "Symbolic model checking using sat procedures instead of bdds," *Design Automation Conference, 1999. Proceedings. 36th*, pp. 317–320, 1999.
15. E. Dubrova, "Random multiple-valued networks: Theory and applications," in *ISMVL*, pp. 27–33, 2006.
16. A. Garg, L. Mendoza, I. Xenarios, and G. DeMicheli, "Modeling of multiple valued gene regulatory networks," *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*, pp. 1398–1404, Aug. 2007.
17. N. Een and N. Sorensson, "Minisat - a sat solver with conflict-clause minimization," in *SAT'2005*, 2005. poster.
18. Chaos, Alvaro, Aldana, Max, Espinosa-Soto, Carlos, Leon, Berenice, Arroyo, Adriana, Alvarez-Buylla, and Elena, "From genes to flower patterns and evolution: Dynamic models of gene regulatory networks," *Journal of Plant Growth Regulation*, vol. 25, pp. 278–289, December 2006.
19. F. Li, T. Long, Y. Lu, Q. Ouyang, and C. Tang, "The yeast cell-cycle network is robustly designed," *PNAS*, vol. 101, pp. 4781–4786, April 2004.
20. R. Albert and H. G. Othmer, "The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in drosophila melanogaster," *J Theor Biol*, vol. 223, pp. 1–18, 2003.
21. M. I. Davidich and S. Bornholdt, "Boolean network model predicts cell cycle sequence of fission yeast.," *PLoS ONE*, vol. 3, no. 2, 2008.
22. A. Faure, A. Naldi, C. Chaouiya, and D. Thieffry, "Dynamical analysis of a generic boolean model for the control of the mammalian cell cycle," *Bioinformatics*, vol. 22, pp. e124–e131(1), 15 July 2006.
23. S. Klamt, J. Saez-Rodriguez, J. A. Lindquist, L. Simeoni, and E. D. Gilles, "A methodology for the structural and functional analysis of signaling and regulatory networks.," *BMC Bioinformatics*, vol. 7, February 2006.
24. L. Mendoza and I. Xenarios, "A method for the generation of standardized qualitative dynamical systems of regulatory networks," *Theoretical Biology and Medical Modelling*, vol. 3, pp. 13+, March 2006.