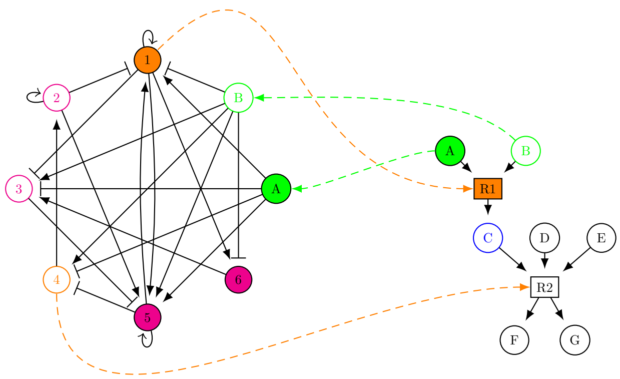
Abstract

Boolean networks have been successfully applied to model gene regulatory networks. Inspired by [**?**] we started to couple boolean with metabolic networks to observe their evolution. To do that, we numerically implemented a fixed size population of organisms which divide upon accumulating biomass. This is achieved depending on their correct switching of reactions and the consequent production of a target molecule, similar to a percolation in the metabolism of that organism. A biomass penalty proportional to the number of enzymes being produced is applied to them, in order to avoid the trivial solution (all reactions on). Each organism has its own boolean network and whenever it divides it produces an exact copy and a mutant one. The food molecules in the metabolic network are also present as ingoing nodes in the boolean network, acting thus as sensors of a varying environment. Some boolean nodes represent enzymes in the metabolism and turns the corresponding reaction on. This effectively couples both networks of each individual. The topology of the metabolic network is shared by all individuals of a population, and different topologies are being proposed as different tasks for the populations to solve.

The main idea consists of observing how the organisms in a population evolve their gene regulatory structure according to the availability of nutrients. To achieve that, we constructed organisms that couple the metabolic network with their boolean network. We couple these two types of networks in the intuitive way: whenever a gene i is switched on, if there is a corresponding chemical reaction in which it works as an enzyme it will promote that reaction in case the educts are all present within the cell. And to sense the presence of some metabolites we can also send the signal from the metabolic network to the boolean network. These nodes in the boolean network that work as sensors are not genes, we will call them sensor nodes.

The boolean network is composed by G genes and S sensor nodes that may have only two states, $\sigma_i \in \{0, 1\}$ and a certain interaction topology. Every gene or sensor node may potentially repress or enhance other genes, and we define a weight $w_{ij} \in \{-1, 0, 1\}$ for the efect that gene or sensor node i has on gene j. This is usually not symmetric. The metabolic network is a bipartite directed network composed by reaction nodes and chemical species nodes.

The figure below shows a general scheme for one possible instance of an organism.



Note that there are no ingoing edges to the sensor nodes A or B, represented in green. We are also distinguishing two types of genes, those that control directly a chemical reaction, represented in orange, and intermediate ones that have only an indirect effect, represented in magenta. The weights of the interaction (repressing or enhancing) are depicted with different arrowhead types. And finally, if a node's state is on, it is filled with solid colour.

The updating mechanism of the boolean network is synchronous and based on threshold functions. Every gene will be updated at the same time, and the state of gene i in the next time step depends on the sum of its inputs multiplied by the corresponding weights according to equation 1. In case it exceeds the threshold θ_i for that gene it will switch on.

$$\sigma_i^{t+1} = \Theta(\sum_{j=1}^{G+S} \omega_{ji}\sigma_j^t - \theta_i), \text{ for } i \in \{1, 2, \dots G\}. \tag{1}$$

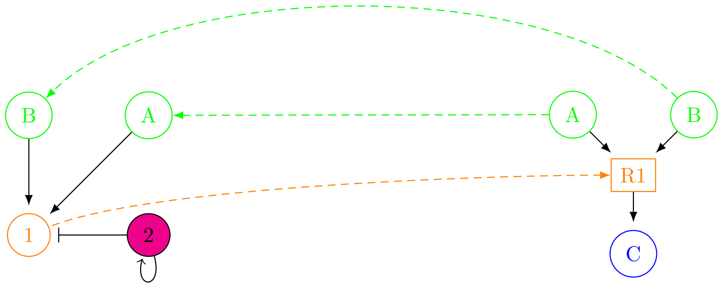
where

$$\Theta(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

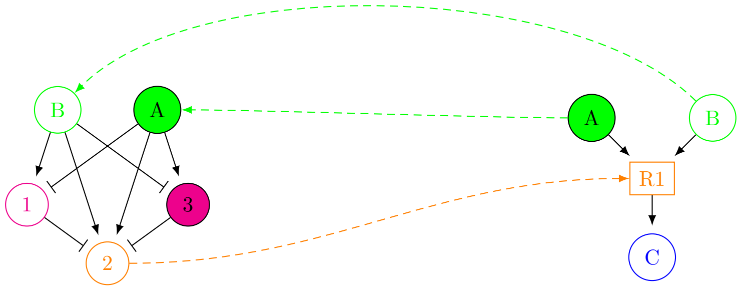
is the Heaviside function.

The population has a fixed maximum size N, and the organisms divide whenever they accumulate a certain amount of biomass. Division leads to mutations in the boolean network structure and eventually the population will adapt to their environment. The environment changes according to some pattern, and most of the time we studied a periodic pattern of availability of nutrients (A and B in this case). Biomass increases when certain target molecules are being produced (and the correct chemical reactions are switched on), and is discounted as a penalty proportionally to the number of genes switched on. This reproduces partially what we expect from biological systems.

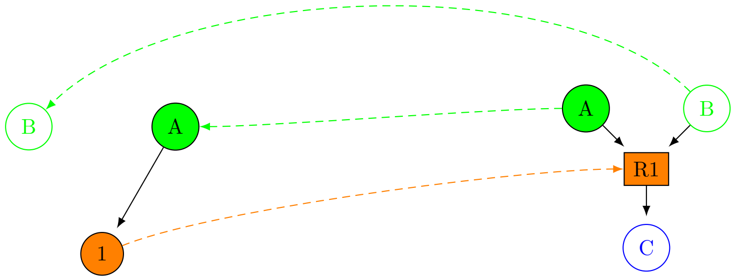
Now let us focus our attention in the (almost) minimal case, where there is only one chemical reaction, with two educts and one product. We assume that the environment changes periodically, such that in only $\frac{1}{4}$ of the time both molecules are present, in $\frac{1}{2}$ of the time only one of them is present and in the other quarter none of them is present. In that case, the organisms have to solve the problem of detecting the simultaneous presence of both educts A and B. So, in a sense, the population is trying to learn an AND function. If we set the thresholds of all genes to 0, we would have the following solutions (of course increasing the number of genes the number of solutions grows as well). The first one would be maintaining one intermediate node always on, as shown below. This circuit works as an AND function.



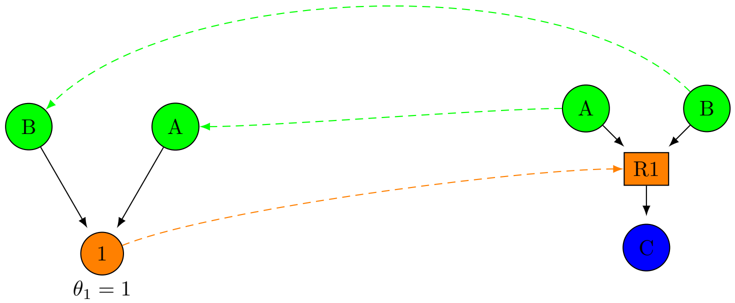
However, this solution keeps always at least one gene working. This means that this solution is worse than producing all the time the enzyme that catalyses the reaction (the trivial solution). Another solution is shown below.



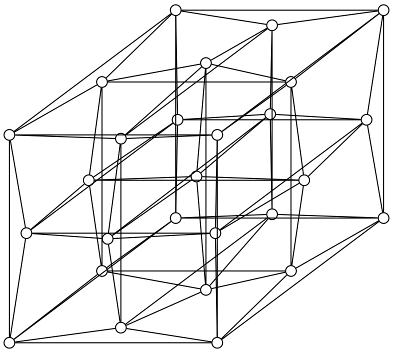
This solution is better and topologically more complicated, but still not good enough. It is better than the trivial solution, but less advantageous than responding to only one of the nutrients. It is important to mention that the fitness of an organism is not explicitly defined in this model. Organisms simply accumulate biomass and divide whenever they reach a certain division threshold. These two examples have similar activation pattern for the enzyme but different division rates (and fitness). The best solution, fixing all the thresholds to zero, would be to produce the enzyme whenever one of the two nutrients - say A - is present. In that case, the organism learns how to shut down its metabolism half of the time. But still, the enzyme is necessary only half of the time it is being produced.



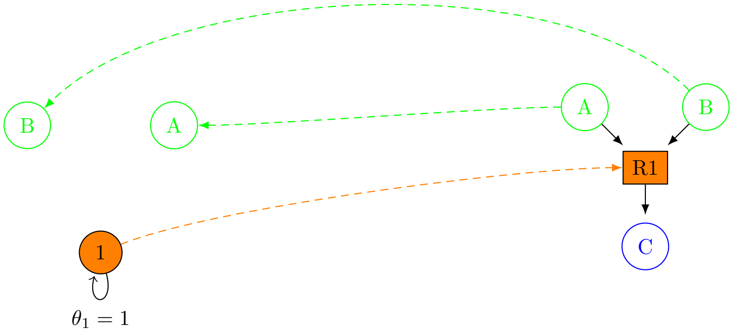
That leads us to investigate the natural problem where each gene has a different threshold, and these thresholds are also subject to mutation. Doing that, learning becomes possible. The boolean network below only switches on the enzyme when both A and B are present, and no other genes are necessary to solve the task.



If we increase the number of educts, the same solution still holds. That justifies the use of varying thresholds. Now the problem becomes understanding how a population would behave inside of a landscape of an implicitly defined fitness. However, given an environmental pattern we can deduce average rates of production of biomass and consequently derive the division rates associated to each genotype. For fixed threshold the genotype network looks like a Hamming graph like the one below for two sensor nodes and one gene. Otherwise, it is the cartesian product between this hamming graph and a path graph.



Given that, we are studying the behaviour of the population on top of this landscape depending on some parameters and conditions. We use thresholds varying from -5 to 4. Then, we can gather all the genotypes into groups of phenotypes and explore how the innovations take place from one neutral network to another. The usual initial conditions are like indicated below.



It is possible to group, in this case, 270 different genotypes into only 8 different fitness classes. To our purposes, only four of them will matter. The first (and less fit) correspond to organisms that never produce the enzyme (SD). The second contains organisms that keep their gene always on (SL), and never interrupt the production of the enzyme, even if the chemical reaction is not possible. The third class (VL3) gathers genotypes that switch off partially when A and B are not present at the same time. And finally, the most fit which correspond to a single genotype that performs the AND function and switches off whenever unnecessary (VL4). We observe the relative amounts of the phenotypes in the population as a function of time. We compared both discrete stochastic simulations with a continuous linear approach. Both capture a transition to extinction and also the fact that, before the population finds the optimal solution it usually goes through an intermediate phenotype.

The green curve corresponds to the VL3 intermediate population before the solution is found. For small mutation rates, the dynamics of the discrete case are governed by rare events, abrupt transitions from one neutral network to the other.

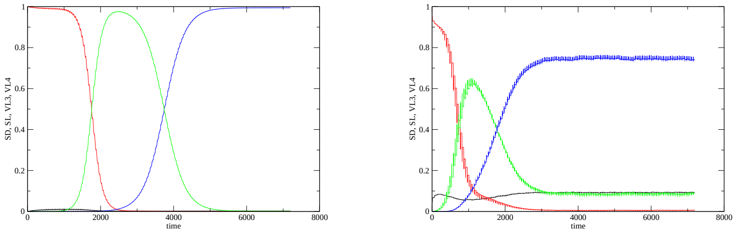


Figure 1: The first row shows the solution of the linear system given by the landscape, whereas the bottom row is an ensemble average of stochastic discrete simulations.

Another important aspect to mention is that, if we increase the period where both nutrients are available, it is not anymore a big advantage to have such boolean networks, and the population unlearns this solution. This is captured by both discrete and continuous approaches. This is what makes the definition of fitness subtle and not simple to deal with. Especially if one thinks of more complex patterns of availability of nutrients. This process is actually governing the evolution of the gene regulatory networks. Another counterintuitive result relates to the division mechanism of the population. We compared two different process - one involving generation overlap and the other without - and they have shown to be considerably different in respect to trajectories in configuration space and its attractors.

We could see that different reproduction mechanisms lead to different attractor structures in the landscape, and increasing the size of the network will also make the process much richer. The next steps go in the direction of increasing the size of both networks and understanding how the interaction between them will select topologies. Being a history-dependent process, evolution is a highly non-ergodic process and unraveling its mechanisms requires a more subtle understanding of out-of-equilibrium statistics. For future projects we intend to use a more detailed description of the metabolic network, involving reaction rates and concentrations of the metabolites, in an attempt to connect boolean networks and flux-balance analysis.