Modelling Evolution of Regulatory Networks in Artificial Bacteria

Yolanda Sanchez-Dehesa, David P. Parsons, Jose Maria Pena, Guillaume Beslon

Math. Model. Nat. Phenom.
Vol. 3, No. 2, 2008, pp. 27-66

Modelling Evolution of Regulatory Networks in Artificial Bacteria

Y. Sanchez-Dehesa[a,c], D. Parsons[a], J.M. Peña[b], and G. Beslon[1,a,c]

[a] LIRIS CNRS UMR5205, INSA-Lyon, Université de Lyon, 69621 Villeurbanne, France
[b] DATSI, Universidad Politécnica de Madrid, 28660 Madrid, Spain
[c] Institut Rhône-Alpin des Systèmes Complexes (IXXI), Lyon, France

Abstract. Studying the evolutive and adaptative mechanisms of prokaryotes is a complicated task. As these mechanisms cannot be easily studied "in vivo", it is necessary to consider other methods. We have therefore developed the RAevol model, a model designed to study the evolution of bacteria and their adaptation to the environment. Our model simulates the evolution of a population of artificial bacteria in a changing environment, providing us with an insight into the strategies that digital organisms develop to adapt to new conditions.

In this paper we describe the principles and architecture of the model, focusing on the mechanisms of the regulatory networks of artificial organisms. Experiments were conducted on populations of artificial bacteria under conditions of stress. We study the ways in which organisms adapt to environmental changes and examine the strategies they adopt. An analysis of these adaptation strategies is presented and a brief overview was proposed concerning the patterns and topological characteristics of the evolved regulatory networks.

Key words: evolution, regulatory networks, modelling, motifs, adaptation mechanisms
AMS subject classification: 9204, 92D10, 92D15

1 Introduction

Prokaryote organisms are very diverse, living in different environments and developing various abilities. Bacteria are found in every ecosystem – some being colonized only by microorganisms – illustrating the impressive adaptation capabilities of prokaryotes. They can be

[1] Corresponding author. Email: guillaume.beslon@liris.cnrs.fr

Article available at http://www.mmnp-journal.org or http://dx.doi.org/10.1051/mmnp:2008054

found, for example, surviving anaerobically in acid elements, in symbiosis with other organisms (e.g., *Buchnera aphidicola*, which lives in symbiosis with aphids, providing essential amino acids for their host), or even in the human intestine where *Escherichia coli* favors digestion and absorption of nutrients.

Bacteria are good examples of organism adaptation. They are able to react to variations in their environment at different levels: bacteria strains can adapt to major environmental changes by a darwinian evolutionary process and individual bacteria can adapt to short-term changes in their environment. To achieve this kind of adaptation at different levels, bacteria have developed a large repertoire of strategies that may themselves be optimized depending on the characteristics of the environment: stability, periodicity, stochasticity, competition...

Although a lot of different strategies (e.g. evolution, regulation, bet-hedging, adaptive mutation, gene amplification, Baldwin effect) have been identified and are relatively well characterized individually, we only have a very partial insight into how they combine with one another: in an idealized environment, one can identify the optimal strategy and mathematically find the optimal parameters. However real environments are far from ideal and there will generally be a wide range of viable adaptation strategies, combining e.g., regulation and evolution, evolution and bet-hedging, regulation and gene amplification or any combination of these. For instance, if the environment changes slowly, bacteria may have enough time to mutate and darwinian evolution can be sufficient to adapt to new conditions. But, they may not be able to conserve complex regulation strategies since mutations quickly degrade regulation mechanisms when these are inactive [14]. Now, if the environment varies a little faster, evolution can be less efficient than regulation, provided that bacteria are able to sense their environment at an acceptable cost and that environmental changes show some regularities (e.g., switches between two different energy sources as in the well-known *lac* operon). On the contrary, rare but unpredictable events put organisms under stress and are known to promote specific adaptive strategies such as the development of mutator strains [44]. All these different strategies imply plasticity at different levels: genetic, metabolic, physiologic, phenotypic, all of which are involved in complex interactions.

These adaptation mechanisms help bacteria to adapt to changing environments. However each has its own tempo, ranging from slow (i.e., darwinian strategy) to fast (i.e. stochastic perturbations leading to phenotypic variability). In the middle, genetic regulation enables a fast dynamic adaptation, enabling cells to react to chemical signals. Regulation is the main mechanism to provide adaptive behavior at a metabolic level. However, regulation never acts alone, it is obviously combined with evolution: genetic variations, gene duplication, gene loss or chromosomal alterations [19] constitute a vast repertoire of variations that can be used by a bacterial strain to adapt to its environment, but that can also provide bacteria individuals with tools to develop more complex adaptation mechanisms. In specific conditions evolution gives rise to regulatory systems that enable fast adaptation to rapidly changing environments. In the case of the *lac operon*, regulation enables the organism to save energy when several food sources are available. It is supposed that regulation is a result of adaptation to changing environments. Yet, it can be shown that such a system can be very sensitive to changes in the environment conditions: Dekel [14] has shown that only a few hundred generations

are necessary for *E. coli* to drastically change its *lac operon* behavior when placed in new conditions. At the other end of the time scale, the *lac* operon is known to have a stochastic behavior [11, 17] and it can be shown that stochasticity of transcription interacts with the regulatory activity of the operon, delaying the operon switch [23]. Thus, while regulation activity has long been supposed to be independent of slow evolutionary changes or fast stochastic variations, it is becoming more and more clear that the interactions of all these adaptation strategies must be studied to fully understand their behavior [22].

It is still a matter of debate in what kind of situation/environment evolution promotes the emergence of regulatory processes and how regulation interacts with the evolutionary process itself. Hypotheses cannot be easily studied on real living systems. Although experimental evolution is possible with micro-organisms [16], tracking changes in genomes, regulatory networks and even phenotypes is almost impossible in "in vivo" tests. An alternative is to use digital organisms to study the genetic bases of adaptation "in silico" [2]. In such artificial models, organisms (i.e., computational data structures) are placed in a synthetic environment that provides them with resources. In this environment the organisms reproduce, mutate and compete for the resources, thus resulting in darwinian evolution. Since the organisms as well as the environment are artificially defined they can both be perfectly and completely described [38]. Such models have already shown their usefulness in studying evolution of robustness [47] or in identifying indirect mutational pressure that regulates genome size [29]. Yet, since most of these models focus on mutational adaptation, they cannot be used to study complex interactions between the different adaptation mechanisms.

The definition of a suitable model to describe this biological process would be useful to tackle many open questions in the literature of this domain: How do organisms adapt to environmental changes? What is the origin of regulatory networks? Why do regulatory networks appear during evolution? How do networks evolve over time? Studying the inclusion of new nodes in already existing regulatory networks and studying the development of new regulatory networks could help to answer some of these questions and provide us with a better understanding of network evolution.

Genetic networks appear to be highly organized: they are modular [21], scale-free [7] and some motifs are overrepresented [4]. Yet, the precise origin of these structures is not fully understood. In particular, it is quite difficult to distinguish between selective origin (the structure of the network is selected because it ensures a correct function in the organism's environment), mutational origin (the mutational process tends to favor some structures, as in the preferential attachment model [7]) and indirect selective origin (the network structure is selected because it is robust to mutation or, on the opposite, highly adaptable). It has been shown that in some specific conditions, modular structures can be selected in evolved networks [20, 25]. Here again, modelling is an essential tool to tackle such questions.

Structure and dynamics of regulatory networks are at the heart of systems biology. The rapid development of this field has been followed by the development of a very active modelling activity of such networks. As far as evolution of regulatory networks is concerned, the work has been focused on the question of topology evolution [25, 26, 49], evolution of network robustness [3, 12, 42] and evolution of artificial functions [5, 6, 18, 32]. Most of these papers

deal with direct evolution of genetic networks (i.e., in the model the network structure is directly modified by the genetic operators – mutations, crossing-over and rearrangements) or selection of the individuals on the basis of the network properties (e.g., selection of a specific topology or selection of a specific regulation dynamic).

Additionally, many studies have been conducted to understand evolution of regulatory networks from a bioinformatic perspective. Phylogenetic studies and sequence comparison provide a quite precise view of the forces that shape bacteria genomes and influences the evolution of their regulatory networks [35]. Thanks to these studies, it is now clearer that large genomic events such as genomic rearrangement, horizontal gene transfer (HGT) [19, 31] or gene duplication play a key role in the evolution of networks [45] and that the topology of the network is for a large part indirectly shaped by the mutational dynamic [13].

All these approaches focus on a specific force that shape the network topology (e.g., mutational dynamic, selection for function, selection for robustness - either mutational or functional robustness, ...). However, in a real biological regulation network, all these forces are at work simultaneously and the network topology results from a compromise between all the constraints a network and an organism must face. These constraints themselves depend on the environmental properties: in a static environment, selection for functional robustness is important while in a randomly (but slowly) evolving environment, the mutational dynamic and/or evolvability property may be crucial for the organism. Thus, to better understand how the environment modulates the emergence of specific network properties, an integrated model is needed in which the appearance of different network topologies during the evolution depends on the dynamical properties of the environment. Moreover, this model should respect the main lines of organisms' evolution. Organisms should own a genetic sequence that allows a large variety of mutational events, a complex genotype-to-phenotype mapping that includes a proteome level and enables the evolution of a genetic network inside the organism. Thus, it should be stratified from a genomic level (the sequence being directly modified by mutational events while all other organization levels are only indirectly modified depending on the effect of the random mutations) to a phenotype level (the phenotype level being the only one subject to selection while the other organization levels are only indirectly selected depending on their influence on the phenotype). The proteome level must respect the core properties of regulatory networks' evolution: the regulation network is neither directly mutated nor directly selected. The nodes of the networks are the proteins of the organism but the links result from a complex interaction between the organisms proteins and its genomic sequence: each protein may or may not interact with the sequence at specific locations, modifying the transcriptional activity of a promoter and, consequently, the transcription rate of one or many genes. Each gene is then transcribed at a specific rate that depends on the intrinsic properties of its promoter and on the influence of the regulation network (including activation, inhibition and self-regulation - see below). The protein concentration is then governed by the transcription rate and by a degradation term. Moreover, the whole transcription/translation process is highly stochastic and it is now recognized that stochasticity influences the fate of organisms [17].

Following these principles, we have developed the "Regulatory Artificial Evolution" model

(RAevol). In this model, artificial "digital" bacteria evolve in a variable environment. Along their evolution, these bacteria acquire genes and evolve a complex genome, a complex regulation network and an adapted phenotype. On an evolutionary time scale, the best individuals are those which evolve the best mechanisms to face environmental variations. We are then able to understand which of these mechanisms are efficient depending on the environmental conditions. In this paper, we first describe the general principle of regulation in prokaryotes and we expose the mechanisms that constitute the core of our model (Section 2). Then we precisely describe the RAevol model (Section 3), focusing on the regulation properties. Finally we present a simple artificial evolution experiment that illustrates the main properties of the model (Section 4) and discuss evolutionary scenarii that may be tested with RAevol.

2 Principles of Genetic Regulation in Prokaryotes

The principles of transcription regulation were described in the 60's by Jacob and Monod [24]. Experimenting with *Escherichia coli*, they showed that the transcription rate of a specific genetic sequence depends on at least three factors: its promoter, which is the initial binding sequence of the RNA polymerase, regulation sites (either activators or inhibitors) where some specific proteins can bind, thereafter influencing the transcription process, and external factors such as the concentration of RNA polymerase in the cell. Note that these principles cannot be considered universal: in eukaryotic organisms, the regulation of transcription activity depends on many different mechanisms, including chromatin dynamics.

Contrary to eukaryotes, in which promoters are generally inactive in the absence of transcription factors (initiation complexes are necessary for the transcription to start and a "naked" promoter will be essentially inactive), prokaryotic promoters and RNA polymerase can directly interact with one another. In the absence of regulatory elements, a promoter will have an inherent activity that mainly depends on its quality. When a promoter has a primary sequence very similar to the consensus sequence, RNA-polymerase can easily bind to it. The initiation of transcription will then regularly occur and the intrinsic transcription level will be high (possibly at a maximum level if the promoter has a very good affinity with the polymerase). In this case, the transcription rate will only depend on extrinsic factors such as the RNA polymerase concentration and quality or the transcription elongation speed).

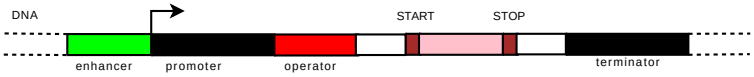
If the promoter affinity to the RNA polymerase is weak, transcription will only rarely be initiated. The quality of the promoter thus determines the transcriptional ground transcription level β (or "basal transcription level", figure 1(a)) [43]. Thus, in the absence of specific regulatory sequences, genes are transcribed at a rate that mainly depends on their promoter strength, maximum transcription rate being bounded by global factors such as the polymerase properties and concentration.

The transcription level can be modified by the action of regulatory proteins. These proteins modify the transcription levels, enhancing or inhibiting gene transcription. In prokaryotes, this process is mainly used to control energy consumption in order to maintain a good balance between food availability and energy, and to adapt to environmental changes.

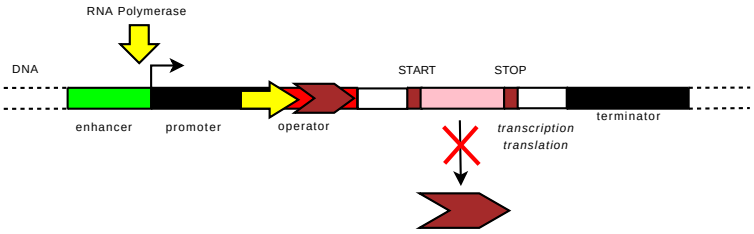
In prokaryotes, inhibition or repression of transcription occurs when a regulatory protein

inhibits the initiation of transcription or the elongation of the transcript (i.e., repressor proteins). Activation of transcription occurs when a protein promotes transcription initiation [48]. When a promoter is activated, its activity can only rise up to a maximum transcription level (meaning that intrinsically efficient promoters can only be marginally enhanced).

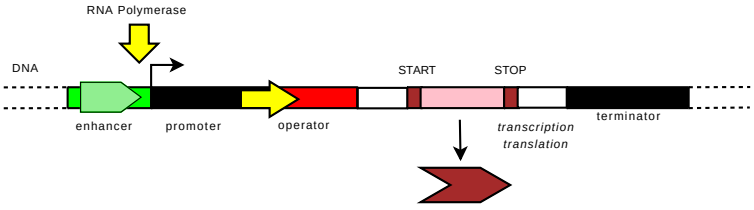
Transcription factors (activation and repression proteins) act by binding to specific regions of the DNA that are near the promoter of the protein they regulate. Repressor proteins bind to a region called *operator* (also called inhibitory region) generally situated downstream from the promoter region. When bound there, a repressor may prevent RNA polymerase from binding or block its displacement along the DNA thus disturbing RNA elongation (figure 1(b)). Activator proteins target *activator-binding sites* are usually located upstream of the promoter region. They promote RNA-polymerase binding, thus enhancing protein production (figure 1(c)).



(a) When no proteins bind the regulatory regions the RNA transcription is done at ground level.



(b) A regulatory protein has targeted the operator. It blocks the polymerase displacement along DNA and prevents it from transcribing the gene. Thus this transcription factor represses the production of the protein associated with this gene.



(c) A protein binds the enhancer region, favoring the RNA-polymerase (top arrow) binding and transcription initiation. Since no inhibitory protein bind the operator, the RNA-Polymerase can transcribe the gene more efficiently, thus enhancing the protein production level.

Figure 1: Transcriptional states in prokaryotes.

In prokaryotes, multiple genes often share a single promoter, its operator and its activator binding sites. These genes are co-transcribed and therefore co-regulated. Such a sequence in

which several genes share their promoter and regulatory regions is called an *operon* because all genes are under the control of a single operator (figure 2).

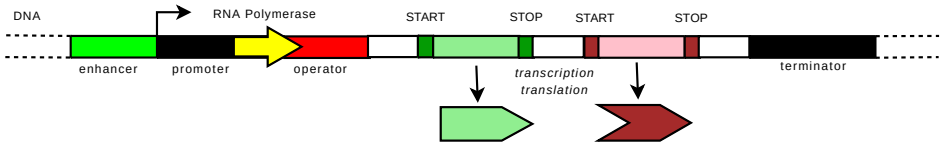


Figure 2: Overview of an operon structure

The best known regulation system is probably the Lactose (*lac*) Operon which controls the lactose-glucose metabolism in *Escherichia coli*. When Monod experimented with the effects of combining sugars as carbon sources for *E. coli*, he found that if glucose and lactose are provided to the bacterium, it first metabolizes glucose and the colony grows fast. When glucose is depleted, the bacteria stop growing. After a short period (lag-phase), bacteria start consuming lactose and the colony grows again. Jacob and Monod later showed that this adaptive behavior comes from a gene regulation mechanism.

In *E. coli*, the lactose metabolism is controlled by an enzyme, the β-galactosidase protein, that breaks down lactose into two simple sugars (galactose and glucose) and by a permease protein that transports lactose from the environment to the cell. The former protein also converts part of the lactose into allolactose.

The β-galactosidase protein is encoded by the LacZ gene and the permease by the LacY gene. Both genes are grouped on an operon structure, the *lac* operon, and are under the influence of the same promoter and the same operator. In fact the *lac* operon contains a third gene, LacA, that encodes for a β-galactosidase transacetylase. A fourth gene, LacI, that is not on the same operon, completes the system by coding for a repressor of the *lac* operon. The repressor protein is able to bind to the *lac* operator, preventing the transcription of the operon (figure 3). However, when lactose is present in the cell, it interacts with the repressor protein, and changes its conformation, preventing it from binding to the *lac* operon. When, the operon is no longer repressed LacY and LacZ can be transcribed. Due to the permease, lactose concentration thus increases, while β-galactosidase is produced and degrades lactose.

The LacI control is an example of negative control. However, it is not sufficient to explain the whole behavior of the *lac* operon. In particular, negative control cannot explain why, in presence of both glucose and lactose, the operon is not transcribed. Indeed, the operon is also controlled by a positive loop: the concentration of glucose is sensed by the cell *via* a signaling molecule, cAMP; the more glucose in the environment, the lower the concentration of cAMP. cAMP binds to an inducer of the operon, the CAP protein, that itself binds on the DNA upstream from the *lac* promoter. Then, the *lac* operon is transcribed if and only if lactose is present in the environment and glucose is not (or no longer) present in the environment[2].

[2]A lots of secondary mechanisms have been discovered. They slightly modify the behavior of the *lac* operon but the two main regulation loops are the negative loop due to LacI and the positive loop due to cAMP binding on CAP (figure 3).

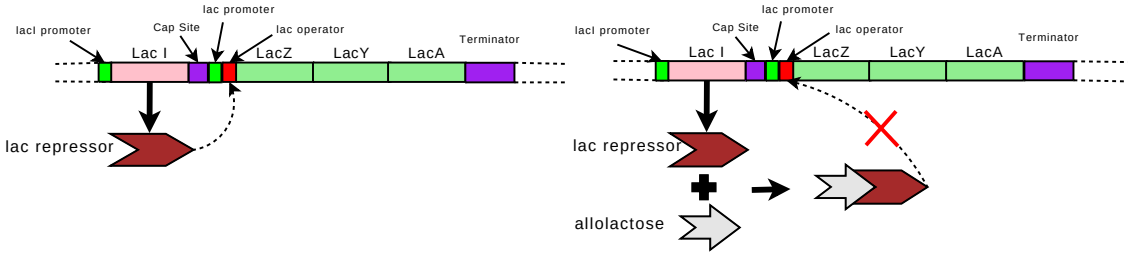


Figure 3: The lac-operon example. When the *LacI repressor* is present (because there is no lactose in the cell), it binds to the operator of the lac-operon, repressing the production of proteins encoded by LacZ (β-galactosidase), LacY (permease) and LacA (β-galactosidase transacetylase). If lactose is present, it is converted into allolactose. Allolactose then binds to LacI, preventing the LacI repressor from binding to the operator. When this occurs β-galactosidase enzyme is produced and degrades the lactose that enters the cell due to the permease enzyme.

At the genome level, all the regulation interactions compose a complex regulatory network. Each network node represents both a gene and the protein it codes for; a link between two nodes means the protein of one node has an influence on gene transcription of the other node (figure 4). Links can be either positive or negative. For example, in figure 4 protein P_3 targets the enhancing region of protein P_2, activating its production. However, when protein P_2 binds to its own operator, it inhibits its own production.

The nature of the transcription network makes its evolution difficult to understand. Since the links represent complex interactions between proteins and specific genetic sequences, they cannot be modified independently: when a genetic sequence varies (e.g., due to point mutation), it perturbs all the interactions between itself and the proteins susceptible to bind to it. Consequently, the influence of the mutations on the network dynamics is a complex process where links are modified collectively. That is why the evolutionary dynamics of regulation networks cannot be fully described by models in which mutations act at a link level (i.e., by adding/deleting single links or changing the weights one by one).

3 Regulation in Artificial Evolution, the RAevol Model

The RAevol model (from Regulatory Aevol Model) is an extension of the "Artificial Evolution" (Aevol) model, developed previously in our team to study robustness and evolvability in organisms [27, 28, 29, 30]. In previous studies, it has been used to demonstrate how individuals adapt their evolutionary strategy to the rate of mutational events. When organisms have low mutation rates, they accumulate non-coding sequences. On the contrary, high mutation rates lead to compact genomes with few and short non-coding sequences. Furthermore, when mutation rates are very high, organisms cannot maintain a large number of genes. Thus, they have to adapt their genome structure to be more robust even though this impairs their capacity to adapt. The Aevol model is well suited for our study because it

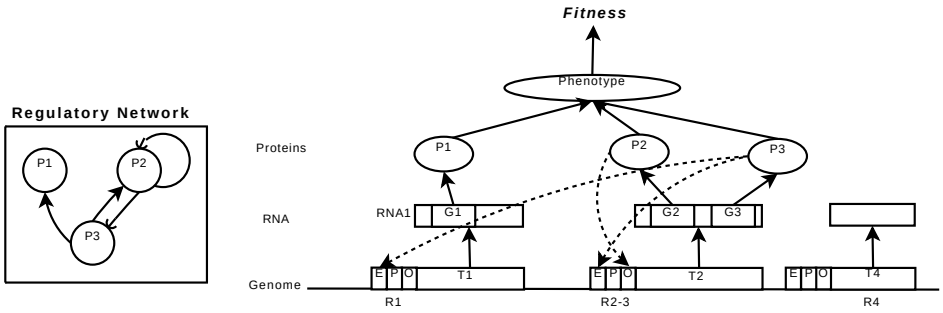


Figure 4: Organization of a regulatory network. A protein P_i can bind to a regulatory region (enhancer or inhibitor site) R_j. There, it regulates the transcription of genes G_j in the T_j region. In this example G_2 and G_3 form an operon that is controlled by the R_{2-3} region. Both genes are co-regulated. They are positively controlled by G_2 and negatively controlled by G_3.

already contains most of the elements needed to study evolution and individuals adaptation.

In Aevol, the genome structure is inspired by prokaryote DNA structure: it is a circular double-strand binary string that contains a variable number of genes separated by non-coding sequences. Each gene is transcribed at a specific rate that depends on the quality of its promoter. Contrary to most artificial evolution models, function of genes do not depend on their position on the genetic sequence. Hence, mutations (including point mutations and genomic rearrangements) can change the genetic sequence as well as the genome structure (e.g., number of genes, operon structure, ...). Finally, the genetic sequence is translated into a set of proteins that interact with one another to produce a phenotype (that can be more or less adapted to its environment).

Although it does not include any regulatory process, Aevol includes all the organization levels needed to design an integrated model of genetic regulation. Its behavior is well characterized and it has been shown to be consistent with bacterial evolution.

3.1 Structure of the RAevol artificial organisms

When designing a model, there is a trade-off between model correctness (regarding the biological objects it aims to describe) and simplicity. In the case of digital genetics, a complete description of prokaryotic biochemistry is impossible. Instead, we will define an abstract, artificial, biochemical framework (an "artificial chemistry" [15]) to be used by the digital organisms to perform metabolic functions. In artificial evolution, the most popular artificial chemistry framework derives from genetic programming. It was proposed by T. Ray in the Tierra program [39] and extended by C. Adami who developed the Avida digital evolution environment [1, 37]. In Avida, organisms' genomes are computer programs written in a simplified assembly language. The computation of organism metabolisms is then straightforward: the assembly language is "simply" executed on a virtual computer with a shared memory.

We argue that Avida's artificial chemistry is in fact too straightforward to study the evolution of regulation networks. In Avida, the genome and the metabolism are structurally equivalent. There is no real transcriptional process and any mutation on the genetic sequence has a direct impact on the organism's metabolism. Avida's chemistry is in fact closer to an RNA-World than a DNA-World (obviously, there is no transcription regulation in RNA-Worlds). Other artificial chemistry frameworks have been proposed and used in digital evolution experiments [10, 21]. However, none of them are able to describe the complex interactions between the genome, the proteome and the phenotype that are mandatory to design an integrated model of genetic networks evolution.

In Aevol (and RAevol), the artificial chemistry is based on a mathematical description of organism metabolism. Each organism is an abstract, virtual entity, represented as a mathematical function, $y = F(x)$, where $x \in \Omega$ represents a specific metabolic function and $y \in [0, 1]$ is the efficacy of the organism for this function (more precisely y is the degree of possibility for the organism to perform this function, see below). Therefore, in our digital world, Ω represents the abstract set of metabolic functions that can be performed by the organisms. To keep the model simple, Ω is a one-dimensional space, i.e., an interval (actually, in all our experiments, $\Omega = [0, 1]$). This means that, in Aevol and RAevol, a metabolic function is described as a real number and that all metabolic functions are topologically organized in Ω meaning that there is a sort of "proximity" (similarity) between metabolic functions. This mathematical description was inspired by fuzzy logic and the theory of possibility [51]. Following the theory of possibility, F is a *possibility distribution*: the space Ω can be seen as the set of metabolic functions that the individual can achieve, and F as the degree of possibility with which a specific function x is achievable by the organisms (a zero possibility meaning that this function will not be performed while a degree 1 means that it will actually be performed). F is formed from the sum of all the metabolic subfunctions accomplished by the protein, by using operators provided by fuzzy logic theory, where each subfunction is described as a fuzzy set.

Fuzzy logic provides a set of boolean operators that enables us to combine the different metabolic functions within an organism (described as fuzzy sets) and to compute the resulting metabolism. Our metabolic chemistry must be complemented with a DNA/RNA translation process. DNA and RNA are sequences that do not directly contribute to the metabolism but can be transcribed and translated into metabolic elements. In our model, the DNA/RNA chemistry is based on binary sequences: DNA is a binary double-strand circular sequence and RNA sequences are described as linear binary sequences.

Most evolutionary models are based upon two-level description of organisms: given a specific phenotype, one has to find an appropriate genetic description and then the genetic operators that can manipulate the genome. In Aevol/RAevol, we introduced a third description level: the proteome. In the model, proteins are the knot that tie all the elements together: genes are sequences that are to be translated into proteins, phenotypes result from proteins interactions, proteins are the nodes of the regulation network, etc. These interactions occur at different levels of description, which implies that proteins will need to be described at these different levels (figure 5):

- From a genetic point of view, a protein can be described as a linear sequence (i.e. primary sequence) translated from a gene thanks to a *genetic code*;
- From a metabolic point of view, proteins contribute to the phenotype of the organism. Each protein is described as an elementary possibility distribution f in Ω whose parameters are deduced from the protein's primary sequence thanks to a *functional code*. In turn, the intensity of the protein's metabolic activity depends on its concentration in the organism.
- From a regulatory point of view, proteins may interact with some specific locations on the genome (namely enhancers and operators), thus modifying the transcription level of genes. A third code will be used to compute the affinity of a given protein with a given regulatory region (*regulatory code*).

We consider that the activity of a protein depends both on its intrinsic capability (i.e. on its primary sequence) and on its concentration in the cell. The concentration is directly modulated by the transcription activity (i.e., by the number of mRNA). Consequently, a cell can modulate its protein production either by gene duplication/deletion or by gene regulation.

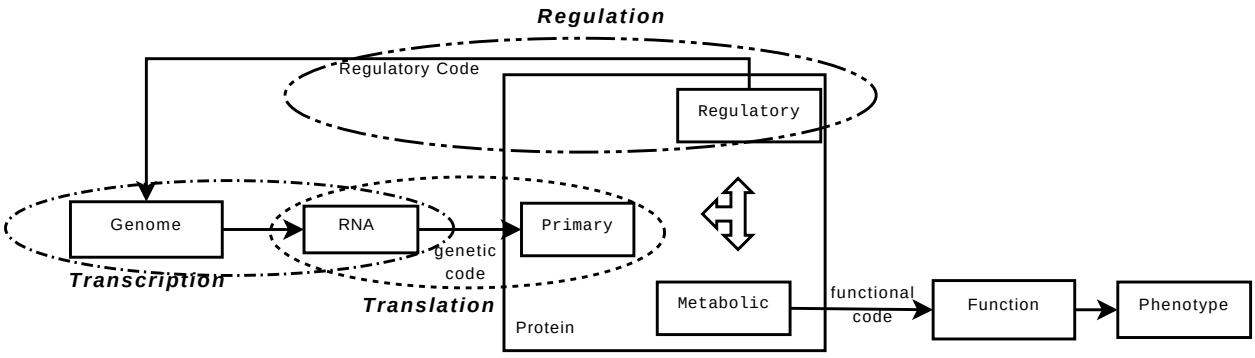


Figure 5: Overview of all the protein roles in the model. Proteins play different roles depending on the elements they interact with. They are translated from the genome (actually from mRNA); they can regulate mRNA transcription in addition to contributing to the phenotype.

In the next section, we will carefully describe the model following the translation process that goes from the genotype to the phenotype (section 3.2). Since the translation process is not strictly linear, we will then describe more precisely the regulation model (section 3.3). Finally, we will describe the global population level in which selection and competition will actually take place (section 3.4).

3.2 From genotype to phenotype in RAevol

The genome is coded as a double-brand binary string, inspired from the bacteria's genome. The first step in genotype-phenotype mapping is genome decoding. The genome sequence

is parsed to identify promoters and genes. Once genes are located they will be transcribed and translated to compute the organism's phenotype.

3.2.1 Transcription: From DNA to mRNA

Both strands of the binary genome are parsed to find the promoter-terminator structures. A consensus sequence was defined and a genome sub-sequence is considered as a promoter when its Hamming distance d from the consensus is less than or equal to the maximum distance d_{max}. In all our experiments, we used 0101011001110010010110 (22 bases pairs) as a consensus sequence and fixed $d_{max} = 4$. The ground transcription state $\beta \in [0, 1]$ (section 2) of the promoter depends on the similarity between the promoter and the consensus sequence (equation 3.1).

$$\beta = 1 - \frac{d}{d_{max} + 1} \tag{3.1}$$

Note that in the model, all concentrations and rates are given in arbitrary unities. Here, the transcription rate is considered to be bound by extrinsic factors such as the concentration and efficiency of the RNA polymerase. The maximum transcription value is the same for all promoters.

The transcription level is modulated by all the protein-genome interactions that take place in the regulatory regions (figure 6). In the model we consider two separate regulation site of 20 nucleotides. The enhancer region (respectively the operator) is situated immediately upstream (resp. downstream) of the promoter. If a protein is able to bind to one of these two regions, it modifies the transcription level of the promoter. Therefore, the actual transcription rate $s_p(t)$ of a promoter depends on its ground state β, on its regulators activity E_{ip} (activation of the promoter p by the i^{th} protein) and I_{ip} (inhibition of the promoter p by the i^{th} protein[3]) and on their concentration $c_i(t)$ at time t. For example, if a transcription factor binds to an enhancer region, it stimulates the production of the associated protein (for a complete description of the regulation model see section 3.3).

Terminator signals are modeled on the stem-loop structure of bacteria ρ-independent terminators. Here any sequence of the form $abcd***\bar{d}\bar{c}\bar{b}\bar{a}$ is considered as a potential terminator (where a, b, c and d are binary nucleotides and $\bar{a}$, $\bar{b}$, $\bar{c}$ and $\bar{d}$ are there complementary bases. $***$ denotes any three nucleotides long sequence). Hence, the transcription is processed downstream from the end of the promoter to the first stem-loop structure found on the sequence. The resulting sequence (mRNA) is an oriented single-strand binary sequence. Notice that a single DNA sequence can be translated several times on the same strand or on the two complementary strands.

[3]In the remainder of the paper, we will use indifferently $s_i(t)$, E_{ij} and I_{ij} to denote the transcription rate (resp. activation and inhibition activity) of promoters and genes. Indeed, we consider that the transcription of a gene is only governed by its promoter.

3.2.2 Translation: From mRNA to protein primary sequence

Once an mRNA has been computed, it is parsed to search for coding regions. Each coding region is then translated into a protein according to an artificial genetic code that associates DNA codons with amino-acids (AA). In the model, there are six amino-acids (see section 3.2.3) so we need eight codons to code for both these AA and the START/STOP codons (there is no redundancy in our genetic code). The translation process is straightforward: the initiation signals are first localized on the mRNA (the initiation signal is the motif $011011***000$, where 011011 represents a Shine-Dalgarno-Like sequence and 000 is the START codon[4]).

From the START codon, the protein sequence is extracted three nucleotides by three nucleotides (i.e., codon by codon) until the termination signal (STOP codon) is found on the same reading frame. Each codon is then translated into the associated amino-acid (table 1). A given mRNA sequence can contain several initiation signals, thus forming an operon structure. One single sequence can in fact code for various genes (and proteins) if several initiation signals are found on different reading frames (genes can also overlap due to the transcription of both strands).

3.2.3 "Folding": from primary sequence to metabolic activity

In this model, a protein contributes to phenotype by its metabolic activity. The metabolic activity is represented as a possibility distribution $f : \Omega \rightarrow \Re^+$ with a standard shape (here f is a piecewise-linear function – actually an isosceles triangle, figure 6). Hence, it can be fully described by three parameters:

- The position of the triangle on the metabolic axis (i.e., its mean $m \in [0, 1]$). This represents the main protein process;
- The height $h \in [-1, 1]$ of the triangle. This determines the maximal possibility degree of the protein (i.e., its activity for its main process). Proteins can either activate ($h > 0$) or inhibit metabolic functions ($h < 0$). The possibility degree of the metabolic contribution is given by $|h|$;
- The half-width $w \in [0, w_{max}]$ of the triangle. This represents the set of metabolic process the protein can contribute to. This parameter expresses the protein pleiotropy (i.e., its ability to achieve different – but related – metabolic processes).

The protein contributes to the set of biological functions ranging from $m - w$ to $m + w$, with a maximal efficiency degree h for the function m. The parameters of the protein are

[4] Although the precision of the model may seem excessive (e.g., Shine-Dalgarno sequence) one has to bear in mind that the model must respect some relative probabilities. Here, the Shine-Dalgarno sequence is used to reduce the probability of initiating the translation process (regarding the probability of finding a STOP codon). Similarly, in section 3.2.1, the complex structure of terminator sequences was used to ensure that terminators are relatively frequent but that no short motifs are excluded from mRNA sequences.

directly computed from the primary sequence of the protein. Once the primary sequence is obtained from the mRNA sequence, three subsequences of codons are extracted according to the metabolic function of each amino-acid (table 1). Each subsequence is then converted into a binary sequence that can be decoded into an integer value (we use the gray code to avoid Hamming-cliffs difficulties). Finally, the three parameters are normalized in the appropriate range depending on the length of the binary sequence, to get the final m, w and h values. Note that a protein can have no metabolic activity if its w or h values are null (degenerated protein). However, this does not mean that it has no influence on the phenotype: a degenerated protein can still have a regulatory influence on the genetic network.

Codon	000	001	010	011	100	101	110	111
Translation function	START	STOP	-	-	-	-	-	-
Amino-Acid	-	-	w_0	w_1	m_0	m_1	h_0	h_1
Metabolic function	-	-	W	W	M	M	H	H
Value	-	-	0	1	0	1	0	1

Table 1: Genetic code in Aevol/RAevol model.

Figure 6 summarizes the overall transcription-translation-folding process. In this example, the mRNA sequence is 100111011101111011010. It is translated into the $m_0h_1w_1m_1h_1w_1w_0$ amino-acid sequence. The three parameters are then given by the three subsequences 01 (M subsequence, length 2), 110 (W subsequence, length 3) and 11 (H subsequence, length 2). Interpreting theses binary sequences with the Gray code we obtain three integer values (1, 3 and 2). Then, these values are converted into real values according to the length of their binary sequence ($\frac{1}{3}$, $\frac{3}{7}$ and $\frac{2}{3}$) and normalized. Finally we get $m = 0.33$ (m is normalized between 0 and 1), $w \simeq 0.02$ (w is normalized between 0 and $w_{max} = \frac{1}{30}$) and $h = 0.33$ (h is normalized between -1 and 1).

3.2.4 Biochemistry: from molecules to phenotype

When a protein i is translated from the genetic sequence, its parameters m_i and w_i are directly issued from its primary sequence. However, at a time t, the actual efficiency $H_i(t)$ of a protein i depends on its intrinsic efficiency h_i modulated by its concentration $c_i(t)$ in the organism (see section 3.3 for the computation of protein concentrations): the higher the concentration, the higher the metabolic activity. This is simply done by using the protein concentration as a scaling factor for the metabolic fuzzy set of the protein ($H_i(t) = |h_i| \cdot c_i(t)$). Then, the actual possibility set to be used for phenotype computation is an isosceles triangle of mean m_i, half-width w_i and height $H_i(t)$.

To compute the phenotype of an organism (i.e. the degree of possibility $F(x)$ with which it performs each function $x \in \Omega$) we must combine the individual actions of each protein. Each protein is represented by a possibility distribution $f_i()$, that can either achieve a set

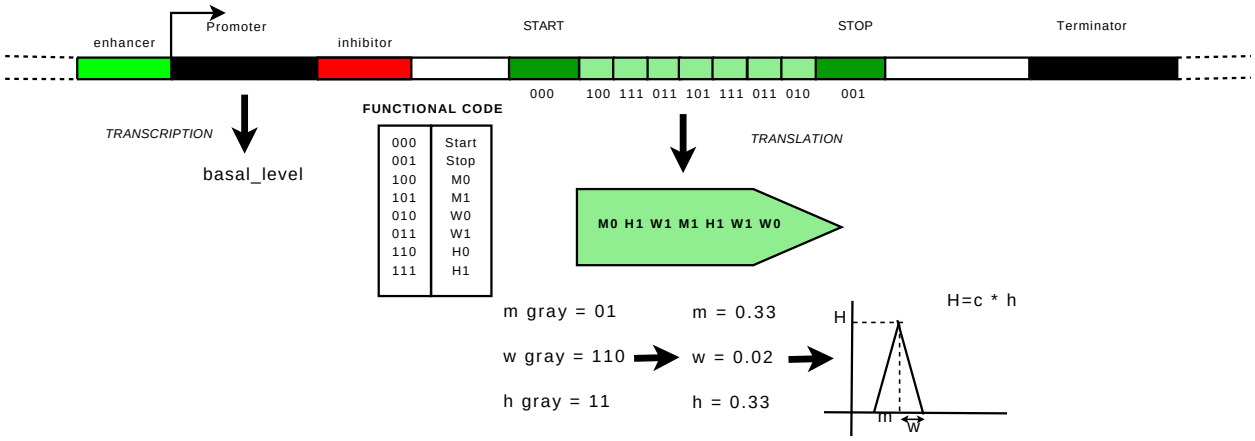


Figure 6: Overview of the transcription-translation-folding process. Once the promoter and the terminator are located, boundaries of genes (START and STOP sequences) are identified and the gene is translated into a protein primary sequence. Three subsequences are then extracted from the primary sequence and decoded to compute the three parameter values that determine the metabolic contribution of the protein. Notice that the exact activity level of the protein (H) depends both on its intrinsic activity (h) and on its concentration (c).

of metabolic processes ($h_i > 0$) or inhibit them ($h_i < 0$). The global functional abilities of an organism are the functions that are activated by at least one protein while not being inhibited by at least one other protein. More formally, we can use boolean operators to compute the phenotype. First of all we compute the activated functions F_A and then the inhibited functions F_I. The activated functions F_A are the functions activated by protein 1 (f_{A_1}) OR by protein 2 (f_{A_2}) OR ... OR by protein n (f_{A_n}). The inhibited functions F_I are calculated following the same procedure as F_A, by using the functions inhibited by protein i (f_{I_i}). Now the global functional possibility distribution F is equal to the combined possibility distributions of all the activated functions F_A AND NOT the possibility distributions of all inhibited functions F_I [28]. In terms of fuzzy sets, this leads to equation 3.2.

$$\boldsymbol{F} = \boldsymbol{F_A} \cap \overline{\boldsymbol{F_I}} = (\cup_i \boldsymbol{f_{A_i}}) \cap \overline{(\cup_j \boldsymbol{f_{I_j}})} \tag{3.2}$$

where $\boldsymbol{F}$ (respect. $\boldsymbol{F_A}$, $\boldsymbol{F_I}$, $\boldsymbol{f_{A_i}}$ and $\boldsymbol{f_{I_j}}$) is the fuzzy set corresponding to the possibility distribution $F()$ (respect. $F_A()$, $F_I()$, $f_{A_i}()$ and $f_{I_j}()$).

To combine proteins possibility distributions, we use the Lukasiewicz fuzzy operators:

$$\begin{cases} \text{NOT} & : & f_{\overline{A_1}}(x) & = & 1 - f_{A_1}(x) \\ \text{OR} & : & f_{A_1 \cup A_2}(x) & = & \min\left(f_{A_1}(x) + f_{A_2}(x), 1\right) \\ \text{AND} & : & f_{A_1 \cap A_2}(x) & = & \max\left(f_{A_1}(x) + f_{A_2}(x) - 1, 0\right) \end{cases} \tag{3.3}$$

Note that in RAevol, the protein concentration can change over time. Thus, all the fuzzy sets must be considered as dynamic functions $f(t)$. However, in the experiments presented in section 4, the global phenotype is computed only once, after a transient period.

3.2.5 Struggle for life: from phenotype to fitness

Our interest in the phenotype of organisms is not the phenotype itself but its adaptation to the environment. In Aevol/RAevol, the environment is modeled as a fuzzy set of functions that are assumed to be useful in this ecosystem. We then define a possibility distribution $E(x)$ that specifies the optimal degree of possibility for each biological function ($E(x)$ can vary over time, either at an evolutionary time scale or at an individual time scale). Then, we use the gap g between this optimal function set and the individual phenotype as a measure of the organism's adaptation to its environment (equation 3.4 and figure 7).

$$g = \int_{\Omega} |E(x) - F(x)|dx = \int_{0}^{1} |E(x) - F(x)|dx \tag{3.4}$$

As shown by figure 7, this measure penalizes the under-realized functions as well as the over-realized ones. Once the gaps of all organisms in the population are calculated, we are able to compute the organism's adaptation and fitness. The adaptation of an organism will then be inversely proportional to the gap (the smaller the gap, the better the adaptation) and the fitness results from a competition with the other organisms in the population. In RAevol, the computation in based on a rank-based selection algorithm: the N organisms are ordered from the least adapted to the best. Then, the reproductive probability P_i of an organism is proportional to its rank r_i in the list. Other selection schemes are also available in the model such as adaptation-proportionate selection or direct exponential-rank-based selection (see [9, 27, 30] for details).

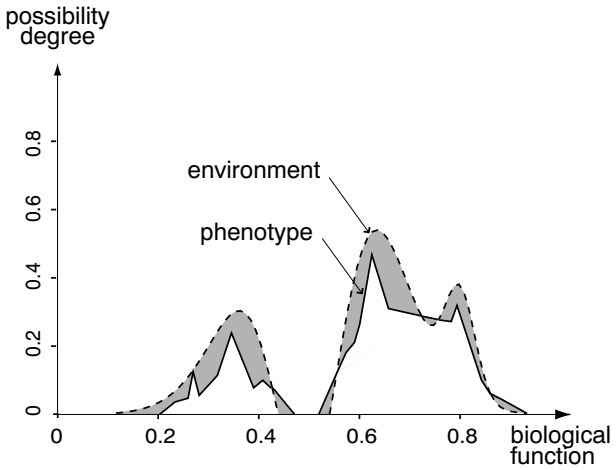


Figure 7: Measure of an individual adaptation. Dashed curve: environmental distribution $E(x)$. Solid curve: phenotypic distribution F (resulting profile after combining all proteins). Filled area: gap g.

3.3 The regulation mechanism in RAevol

The main difference between Aevol and RAevol is the explicit modelling of protein concentration and the modelling of an individual time distinct from the evolutionary time. In RAevol, the proteins are explicitely produced at a given rate that depends on the regulation network and degraded at a constant rate. Their contribution to the metabolism is modulated by their concentration in the cell (section 3.2.4). So, the phenotype of the organisms is no longer a *constant* set of metabolic functions (as it was in Aevol). Now it is a *dynamic* set of functions that can change during the "life" of the individuals.

3.3.1 Computation of proteins concentration

In RAevol the protein concentration depends on three factors: the promoter quality, the degradation rate and the regulation activity. The promoter quality gives the ground transcription state β (equation 3.1, section 3.2.1). The degradation rate is considered constant for all the proteins (exponential decay). Now, the regulation activity depends on all the proteins present in the organism, their concentration and the intensity of their regulatory activity on the operators and on the enhancers.

So the protein concentration $c_i(t)$ is governed by the following equation:

$$\frac{\partial c_i}{\partial t} = s_i(t) - \phi c_i(t) \tag{3.5}$$

where $s_i(t)$ represents the transcription/translation rate of protein i at time t and ϕ the degradation rate (assumed to be constant in the model). The initial concentration of a protein is given by the promoter ground state: $c_i(0) = \beta_i$. We plan to use an initial state $c_i(0)$ equal to the proteins' concentration in the mother cell at the time it divides.

As we have seen in Section 2, the transcription process is regulated by transcription factors that can decrease the transcription rate from the ground state to zero (inhibitors) or increase it up to a maximum value that depends on extrinsic factors (mainly the RNA polymerase). Furthermore, the transcription factors' activity depends on their ability to bind to the DNA molecules at specific locations (enhancers and operators). In RAevol, this regulation process is modeled in two steps: first, we list the regulation capacities of all the proteins on all the promoters (activation and inhibition). This gives us the topology of the regulation network. Then, given the topology and the current concentration of each protein, we are able to compute the regulation activity exerted on each node (i.e., on each gene), and hence to deduce the transcription rate of each protein.

Here, we consider the simplified situation in which the transcription factors activities are purely additive. Therefore, at time t the global activation exerted on the promoter[5] i is given

[5]For sake of simplicity, we consider here the case of a one-to-one association between promoters and genes. Thus, the promoter i is supposed to govern the transcription of the gene i. In the model – and in the real life – the association is not one-to-one, e.g., in case of operon structures.

by:

$$A_i(t) = \sum_j c_j(t) A_{jI} \tag{3.6}$$

where A_{jI} represents the positive regulation activity exerted by the protein j on the promoter I (see next section for the computation of the individual regulation activities). Similarly, the whole inhibition activity is given by the sum of the individual inhibitions modulated by the proteins concentration:

$$I_i(t) = \sum_k c_k(t) I_{kI} \tag{3.7}$$

Then, the transcription activity is given by a Hill-like kinetic [36] scaled in order to respect the basic principles of prokaryotic transcription (see section 2): without any regulators, the promoter is transcribed at the ground state β. It can be up-regulated to a maximum level (that also depends on the strength of the promoter) and down-regulated to zero. The general equation that describes the transcription rate over time is defined as:

$$s_i(t) = \beta_i \cdot \left(\frac{\theta^n}{I_i(t)^n + \theta^n} \right) \left(1 + \left(\frac{1}{\beta_i} - \beta_i \right) \left(\frac{A_i(t)^n}{A_i(t)^n + \theta^n} \right) \right) \tag{3.8}$$

where n and θ are constant coefficients that determine the shape of the Hill-function (in simulations presented in section 4, we used: $n = 4$ and $\theta = 0.5$).

3.3.2 Computation of the binding properties

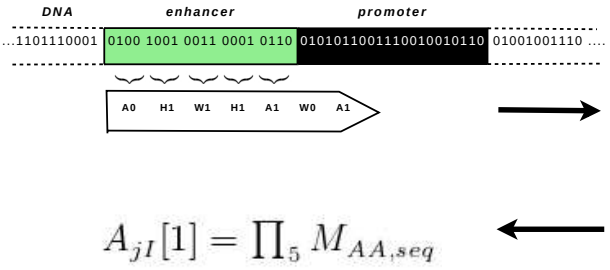
The mechanisms that regulate gene expression in prokaryotes are very diverse and most of them are only slightly characterized. Therefore, a precise modelling of regulation is beyond the scope of a digital evolutionary model. In RAevol, we chose to describe the regulation activity in a simple way: as described in section 2, in a first approximation one can consider that the regulatory property of a transcription factor depends on its ability to bind to the DNA at specific locations (binding sites). Moreover, the contribution of the transcription factor to the promoter activity is strongly dependent on the position of the binding site relative to the promoter.

In the model, each promoter is surrounded by two binding sites of 20 base-pairs (i.e., 20 bits). The upstream site is the enhancer and the downstream site is the operator. Each protein has a probability to bind a given site that depends on its affinity with this site. We will obviously not be able to compute or model a "real" protein-DNA affinity; what we need is a procedure that (i) gives the capacity of any protein to bind to any sequence of 20 bits; (ii) is relatively independent of the metabolic capacity of the protein (i.e. a protein can have a regulatory activity while having no metabolic activity, two proteins with the same metabolic activity can have different regulatory capacities, etc.), (iii) enables us to fix the probability that any protein can work as a transcription factor and (iv) is simple enough to be computed rapidly and therefore to be used in an evolutionary model[6].

[6] In a population of N organisms,having a mean number of genes of M and whose evolution is simulated

To compute the affinity of a protein with a given binding site, we align the primary protein sequence with the binary sequence of the binding site. Since the artificial chemistry of proteins and DNA are not compatible (the "proteome" chemistry is based upon amino-acids – $w_0, w_1, h_0, \ldots$ – and metabolic fuzzy sets while the DNA chemistry is made of bit sequences), the alignments are evaluated thanks to an affinity matrix (figure 8). In this matrix, each cell represents the affinity between a specific animo-acid and a regulatory subsequence of 4 bases. Thus, given the size of the binding site, the affinity will be the maximum alignment value for all possible subsequences of five amino-acids in the protein primary sequence.

For a given protein j and a given binding site I (of protein i), the k possible alignments of the amino-acid sequence on the binding site are computed (e.g. for a protein of length l, $k = l - 4$). For each alignment, we compute the local affinity $A_{jI}[k]$ thanks to the affinity matrix (figure 8). The protein affinity with the enhancer is then given by $A_{jI} = \max_k A_{jI}[k]$.



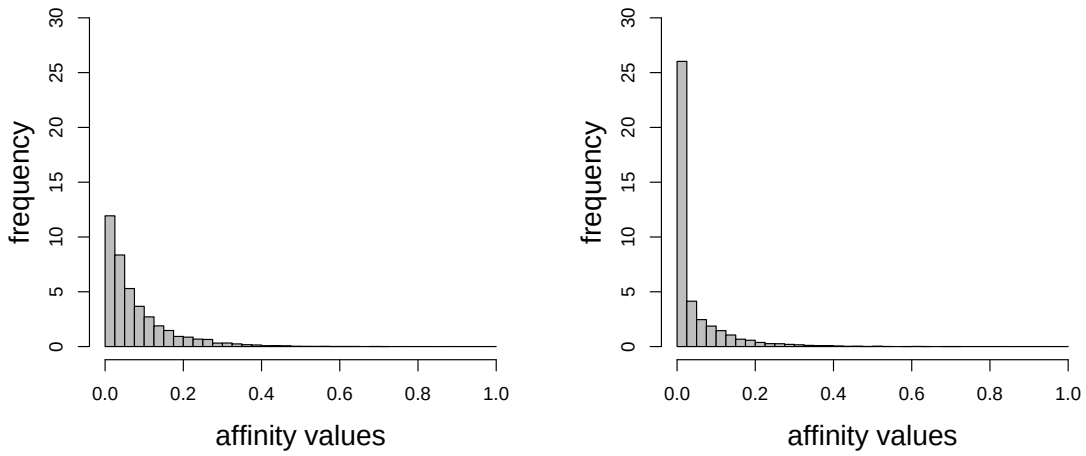
	A_0	A_1	W_0	W_1	H_0	H_1	Start
0000	0.621294	0.797917	0.472641	0.803376	0.049579	0.630847	0.074509
0001	0.749787	0.955064	0.976538	0.483629	0.881366	0.000000	0.731637
0010	0.051620	0.924256	0.010539	0.110479	0.000000	0.466238	0.725215
0011	0.811101	0.511176	0.000000	0.659744	0.493761	0.528889	0.452844
0100	0.000000	0.000000	0.508549	0.094083	0.537400	0.315913	0.877033
0101	0.334670	0.254684	0.268657	0.000000	0.904534	0.853420	0.974473
0110	0.031969	0.516332	0.247085	0.019950	0.386099	0.312712	0.000000
0111	0.387487	0.940186	0.626080	0.754808	0.121314	0.000000	0.000000
1000	0.391194	0.306006	0.326490	0.828298	0.000000	0.357495	0.693670
1001	0.543644	0.000000	0.626592	0.506510	0.136002	0.000000	0.000000
1010	0.295707	0.220555	0.331294	0.000000	0.199195	0.723339	0.000000
1011	0.694573	0.393785	0.828505	0.994704	0.000000	0.339311	0.000000
1100	0.000000	0.623747	0.835268	0.744280	0.935072	0.711471	0.889950
1101	0.114172	0.214537	0.969292	0.402094	0.000000	0.000000	0.453586
1110	0.655020	0.884130	0.794194	0.649743	0.623270	0.000000	0.667005
1111	0.863602	0.015699	0.623551	0.508015	0.000000	0.959038	0.917856

Figure 8: Affinity calculation. In this example protein j is aligned on the enhancer region of promoter I. The first local affinity A_{jI} is computed as the multiplication value of the table entries corresponding to each pair AA/subsequence. We can compute $A_{jI}[1] \simeq 0.0000$. The protein is then shifted leftwards to compute $A_{jI}[2] \simeq 0.01998$ and $A_{jI}[3] \simeq 0.00865$. Then, the affinity of the protein on this enhancer site is given by$A_{jI} = \max_k A_{jI}[k] \simeq 0.0.01998$. This value is to be reported in equation 3.6 to compute the transcription rate of the promoter i.

Using this simple alignment procedure, we are able to define the distribution of regulation by choosing the values in the affinity table. In our experiments, values in the affinity table are randomly chosen following a uniform law between 0 and 1, with the exception of a fixed proportion of cells α that are filled with null values. The parameter α enables us to increase the proportion of null regulation weights (figure 9). Thus we are able to indirectly fix the mean connectivity degree in our networks. Moreover, in RAevol, we actually use two different affinity matrices M_A and M_I. The former is used to compute proteins' affinities with enhancer sites, the latter with operator sites. This allows RAevol users to set different proportions between spontaneous activation and inhibition; experimenters can use either

during T generations, the binding computation procedure will be executed $N*M*T$ times. In the experiments presented section 4, $N = 1000$, $M \simeq 40$ and $T > 20000$.

identical or different matrices depending on whether they want the spontaneous proportion of inhibitory links to be higher or lower than the proportion of activation links or not.



(a) Full matrix

(b) Half-half matrix

Figure 9: Distribution of regulatory links for random binding sites and random protein of 21 Amino-Acids. Left: distribution for a matrix filled with uniform random values ($\alpha = 0$). Right: distribution for a matrix with 50% of null values ($\alpha = 0.5$).

3.4 The Evolutionary Process

In previous sections, we have precisely described the organisms in RAevol. These organisms are subjected to an evolutionary process. In a fixed population, organisms are evaluated thanks to the selection procedure (section 3.2.5). Then, a biased random process is used to determine which of these organisms will reproduce. The reproduction process is based on DNA replication. During this process the DNA can undergo replication errors. These errors (mutations) are governed by operators that are directly inspired from prokaryotic reproduction. Genomes can undergo seven different kinds of mutations: three of them being point mutation, and four large chromosomic rearrangements.

The three point mutations are the switch and the two InDels:

Switch: A randomly chosen nucleotide switches from 0 to 1 or 1 to 0;

Insertion: A random position is chosen in the genome and a small random sequence (1 to 6 bits) is inserted at this point;

Deletion: A random position is chosen in the genome and a small sequence (1 to 6 bits) is deleted at this point.

The four chromosomic rearrangements are the following:

Duplication: Two positions are randomly selected in the chromosome. The segment between these two positions is copied and inserted at a third random position;

Translocation: Two positions are randomly selected in the chromosome. The segment between these two positions is excised and inserted at a third random position;

Large deletion: Two positions are randomly selected in the chromosome. The segment between these two positions is deleted;

Inversion: Two positions are randomly selected in the chromosome. The segment between these two positions is inverted (i.e., the sequences are conserved but they move from one strand to the other).

These mutations affect the genome, and some of them change the genome size (indels, insertions, deletion, duplications and large deletions). Indirectly, they can modify the regulatory network topology by either duplicating/deleting genes or promoter regions. They can modify the affinities between transcription factors and binding regions: when a mutation occurs in the regulatory region of a promoter, the protein's affinities with this region can change. Reciprocally, when a gene undergoes a mutation, the primary sequence of the protein it codes for may change, thus affecting one or both of the protein functions: its regulatory abilities and its metabolic activity.

4 RAevol in action: Scenario and results

The main interest of digital organisms is that they enable practitioners to perform evolutionary experiments on which they have very good control [38, 2]. To make proper use of such models, one has to follow an experimental procedure in which (i) a testing environment is carefully designed, (ii) some parameters of either the environment or the organisms are modified, (iii) the experimenter lets the evolutionary process run for many generations (typically thousands of generations in digital evolution) while carefully gathering information about the evolutionary process and (iv) the experimenter interprets the results as a function of the parametric differences. Thus, although completely artificial, digital evolution is closer to experimental evolution than to mathematical evolutionary models such as population genetic models. It thus makes it possible to test hypotheses that would be out of reach of mathematical models because they cannot sufficiently express the complexity of the system.

In this section, we present a typical experiment with the RAevol model. We will first detail the experimental setup and then compare nine evolutionary experiments (three types of organisms times three different seeds for each one). Finally, we describe the structure of one of the regulatory networks obtained at the end of the evolutionary process.

4.1 Experimental setup

To test the ability of RAevol organisms to develop an efficient regulatory network, we designed a scenario in which, during their lives, the individuals must alternatively achieve two different sets of metabolic functions. In the first set, individuals have to perform three groups of metabolic functions, modeled as three lobes in the Ω space (the exact distribution of possibility of the environment, E_1, is presented on figure 7). When initialized the organisms phenotypes only depend on the basal level of their promoters. After a short transient period (10 simulation time steps), the regulatory networks are very likely to have changed the protein concentrations (see figure 20(b)). It is only at this stage that the organisms are tested for the first time. At time 10, the phenotype is compared to E_1, resulting in the first gap g_1. Then, the environmental reference is changed (removal of the right lobe, environment E_2, figure 10) and a signaling protein is sent to the organisms. This protein (whose sequence is : $h_1 w_0 h_0 m_1 w_0 h_1 m_1 h_0$) has no metabolic function (because it contains no w_1 amino-acid) but is long enough to be able to bind to the DNA and hence have a regulatory activity. We then wait for a second transient period (10 steps) and the phenotype is compared with E_2, resulting in a second gap value g_2. The fitness of the organism is then computed on the basis of the mean gap value $\frac{1}{2}(g_1 + g_2)$. Given the difference between E_1 and E_2, we can approximate that, for an organism without regulation abilities (NULL context, see below), the minimum gap will be given by half of the difference between the two environmental distributions: $g_{min} \simeq 0.011$

According to this scenario, organisms can develop different strategies depending on their ability to tune their transcription levels. The simplest strategy would be to develop strong operators with a high affinity with the signaling protein. If they are associated with the promoters of the proteins in the right side of the metabolic space (proteins with metabolic functions $x \in \Omega$, where x is close to 1), these operators can repress the transcription of these proteins during the second part of the organisms' "life". A more elaborate strategy would be to develop a complex regulation network, e.g., to activate some proteins (possibly without any metabolic function) that will themselves inhibit others. Such a network could accelerate the metabolism response to the signaling protein. Finally, if the organisms do not succeed in developing a regulation network, they can stabilize on the mean value of the metabolic process in order to minimize their metabolic error.

We simulate the evolution of populations of 1000 organisms in this environment for 25000 generations (organisms are initialized with random genomes of 5000 bp each). Each individual dynamic is simulated during 20 time steps in order to compute g_1 and g_2. Then, the selection process is used to determine which organisms will reproduce and how many offsprings they will have. New individuals will replace the old population, with the population size remaining constant. During the mutational process, organisms undergo mutations with a fixed mutation rate of 10^{-5} mutations per base pair (in these experiments, the mutation rate is the same for all types of mutations including point mutations and rearrangements). Finally, we tested three different types of organisms characterized by their affinity matrix M (the same for both activation and inhibition):

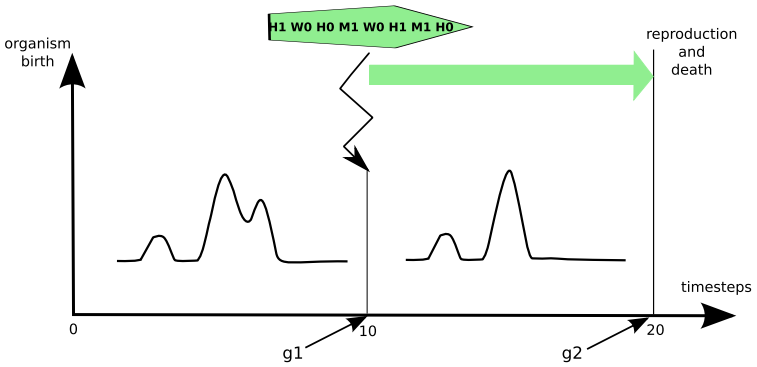


Figure 10: Objective functions to reach during individuals life. In the first stages, three metabolic functions represented by the sum of three Gaussian functions (one being negative). The arrival of an external protein into the cell must be taken into account by the organisms to modify their behavior. The new objective function is a set of two metabolic functions, represented by the sum of two Gaussian functions(one being negative).

NULL: these organisms are used as a reference to test the effect of the regulation process. In the NULL organisms, the affinity matrix M is filled with null values ($\alpha = 1$. So, the NULL organisms are not able to regulate their transcription activity (i.e., the genes are always transcribed at their basal levels).

FULL: in the FULL context, the affinity matrix is initialized with random values in $[0, 1]$ (uniform sampling with $\alpha = 0$). The resulting distribution of regulatory links is shown on figure 9(a).

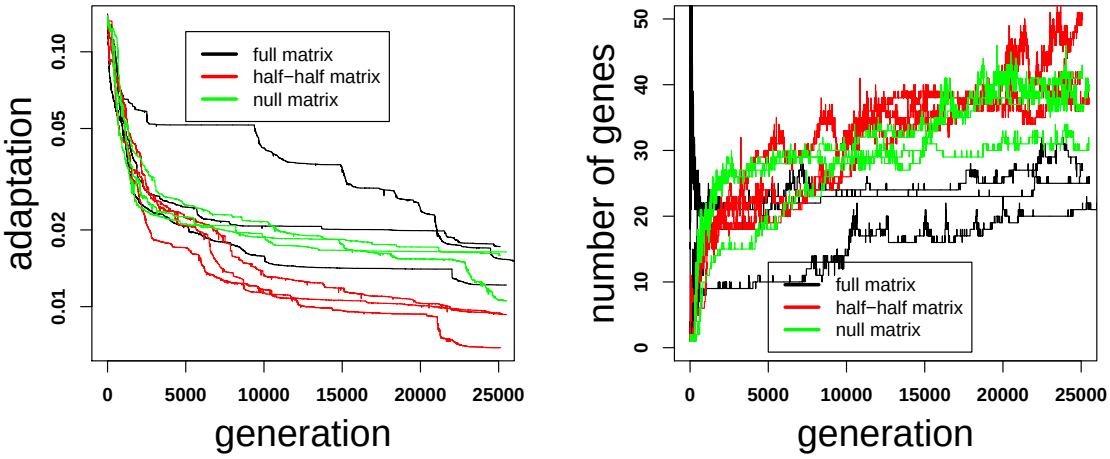
HALF-HALF: in this context, the affinity matrix values are computed in the same way as in the previous one except that half of the entries are filled with a null value ($\alpha = 0.5$). Thus, the affinity values are generally lower than in the second context and a larger proportion of protein/binding sites pairs have a null affinity (figure 9(b)).

For each one of these contexts, we performed three different simulations using three different seeds. Indeed, since the evolutionary process is mainly governed by random events, every experiment must be conducted several times in order to distinguish between the reproducible effect of selection (either direct or indirect) and the effect of drift and contingent events.

4.2 Evolutionary process unfolding

During the 25000 generations of a simulation, the organisms progressively acquire genes that enable them to enhance their metabolic performances (figures 11(a) and 11(b)). During the first generations, organisms acquire "essential genes", i.e., genes with a large metabolic contribution, and hence, the gap g of the organisms quickly decreases. Then, organisms

continue to increase their performances but at a lower speed. During this second phase, organisms adapt to their environment either by acquiring new genes (figure 11(b)) or by optimizing the metabolic contribution of the existing ones. The optimization of the metabolic contribution of existing proteins implies an increase in the average gene length. Indeed as a protein's contribution results from the normalization of the values given by its primary sequence, a greater level of precision can only be achieved by an increase in the length of this sequence: in the model protein parameters, m, w and h, need longer sequences to be more precise (see transcription process in Section 3.2.3). It is worth noting that, in RAevol, as in Aevol, genes are acquired thanks to a duplication-divergence process [27, 29].



(a) Gap g of the best organism. g represents the adaptation level of organisms. Values are represented in log scale.

(b) Number of genes

Figure 11: Adaptation values for the best individual for the three contexts (three seeds for each context). Adaptation value is the gap between the objective function and the metabolic function achieved by organisms (i.e. the reverse of fitness).

4.2.1 Evolution of the genetic structure

The only difference between the types of organisms tested in our experiments is the proportion α of non-null values of the affinity matrix, which ranges from zero (NULL context) to 1 (FULL context). Analysis of different genomic characteristics (genome size, number of genes, mean gene length) and the main phenotype parameter (the gap) clearly shows that the density of the affinity matrix has a strong influence on the course of evolution. Surprisingly, the worst organisms are not the NULL ones (i.e., organisms that are not able to

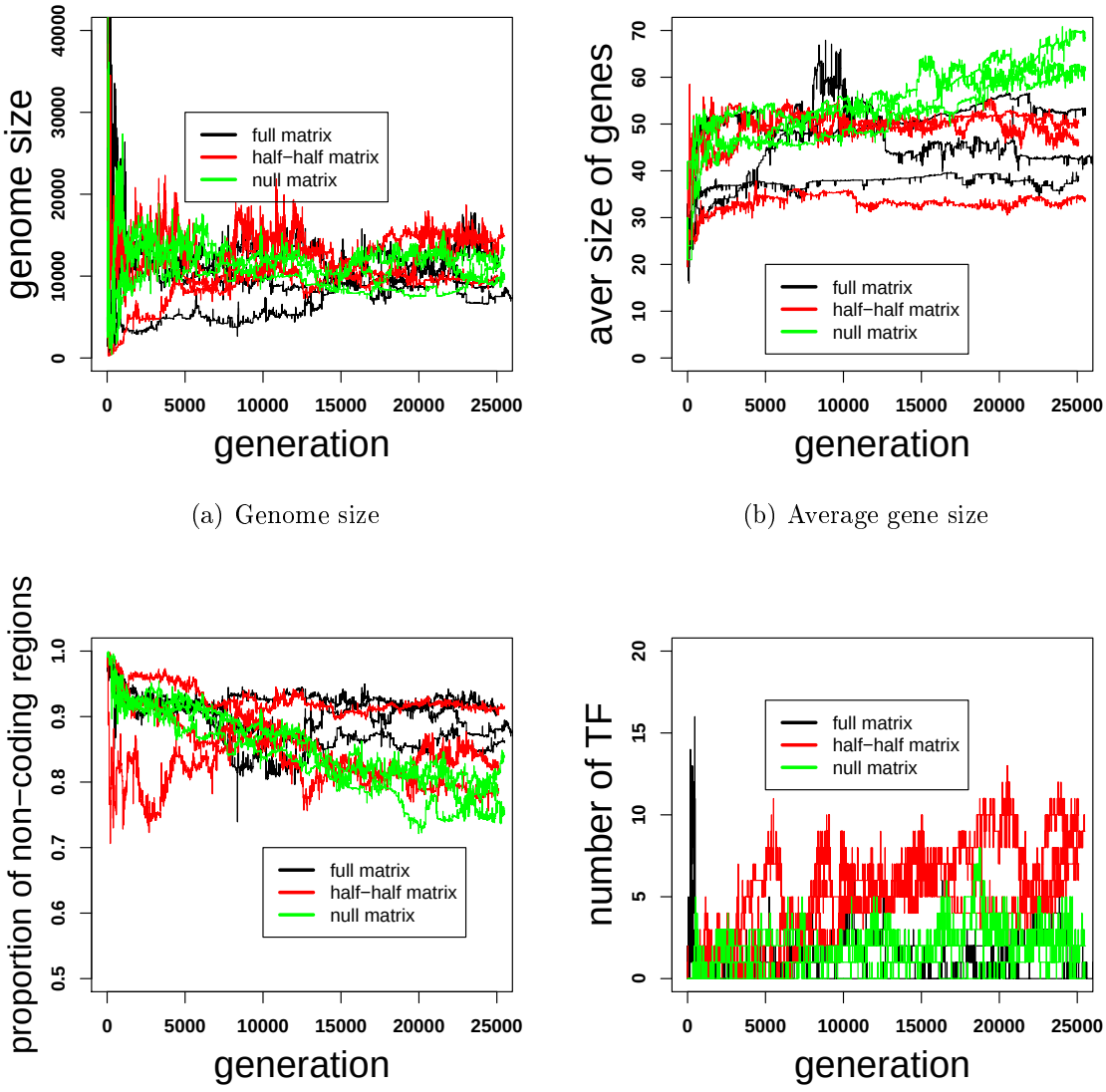
regulate their gene transcription) but the FULL ones (figure 11(a)). This can be easily understood when looking at the evolution of the genetic structure (figure 11(b)): in the FULL context, the genomes contain fewer genes than in the two other contexts. In a previous experiment conducted with a simplified version of the model, we have already shown that, in the FULL context, the individuals have a poor evolvability due to the over-connectivity of the regulation network [40, 41]. The high density of the affinity matrix results in a highly connected regulation network (figure 9(a)). Any perturbation of a protein and/or binding site has a high impact on the organism's phenotype (because it systematically affects several genes). Moreover the metabolism and the genetic network are strongly linked, making the equilibrium between them very unstable and thus lowering the organisms' evolvability. It is worth noting that this effect would not be visible in classical evolutionary models of regulation networks because, in these models, the mutations act directly on the regulatory links allowing the organisms to remain evolvable by providing them with the possibility to modify the regulatory links independently of one another.

When looking at the genome size we cannot conclude that the density of the affinity matrix influences the genome size (figure 12(a)). However, if we compare the number of genes and the mean size of the genes in the different contexts we can see that FULL organisms are less compact, having more non-coding regions (figure 12(c)). Using the Aevol model it has been previously shown that these parameters directly depend on the mutational robustness of the organisms [30, 29]. Therefore, we now need to test the robustness of the evolved organisms by artificial mutagenesis experiments[7].

While the FULL organisms are the worst ones, the best ones are not the NULL ones but the HALF-HALF ones. It seems that the mid-density of the affinity matrix gives the regulatory network the ability to evolve in a relatively independent way. While in the two other contexts (FULL and NULL) the number of regulatory links is either null or directly determined by the number of genes (roughly equal to the square of the number of genes), which means the gene network is either fully connected or not connected at all, in the HALF-HALF context the regulatory network is only partially connected. This provides a greater degree of freedom for the organisms to evolve their regulatory network. Figure 13 shows that, in the HALF-HALF context, the number of links evolves continuously while, in the FULL context, it undergoes long stationary phases, resulting in long period of stasis in the organism's fitness.

As we can see in figure 12(d), the number of non-metabolic proteins increases over time. These proteins cannot achieve metabolic functions but they are able to develop regulatory tasks: they can bind to regulatory regions and modify the transcription of associated proteins. They can be considered as transcription factors (TF). Note that TFs mainly appear in the HALF-HALF context. The acquisition of transcription factors is one of the signs that indicate the creation of a complex regulatory network.

[7] In these experiments, an organism is submitted to a repeated mutagenesis process in order to measure the fitness loss.



(a) Genome size

(b) Average gene size

(c) Proportion of non-coding regions

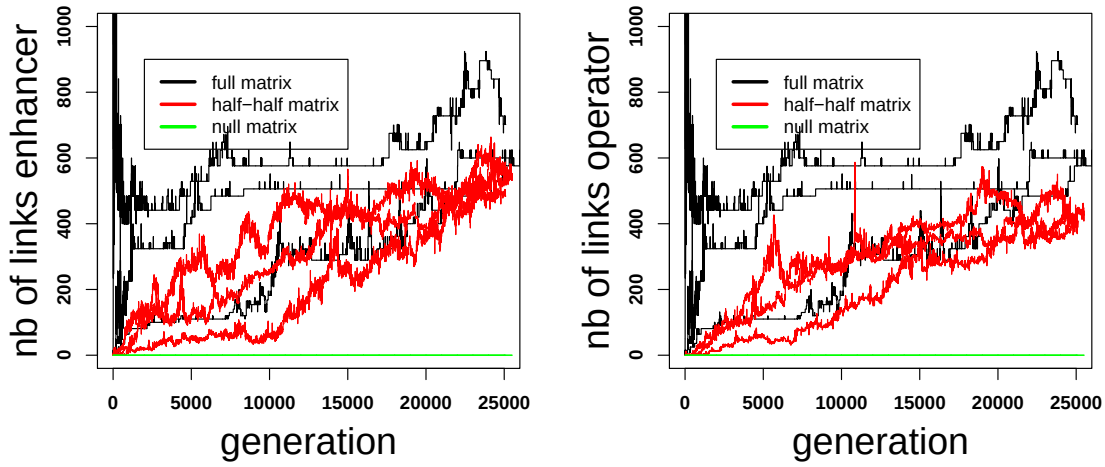
(d) Number of transcription factors (TF, i.e. non-metabolic genes)

Figure 12: Evolution of the genome structure for the best individual of the population. The non coding regions are considered as the genomic sequences between a terminator and the next promoter.

4.2.2 Evolution of the regulation network

Figures 13, 14, 15 and 16 present a global overview of the evolution of the genetic network. While figure 13 shows that links are regularly added to the network (mainly thanks to a

gene duplication divergence process), either the mean link weight (figure 14) or the link weight histograms (figures 15 and 16) are mainly stable. Moreover, in the case of FULL organisms, the link histograms are close to random distributions (figure 15 left columns), showing that, in such conditions, the link weights are mainly contingent. In the case of HALF-HALF organisms, distributions are biased toward null values (figure 16), with a few strong links.



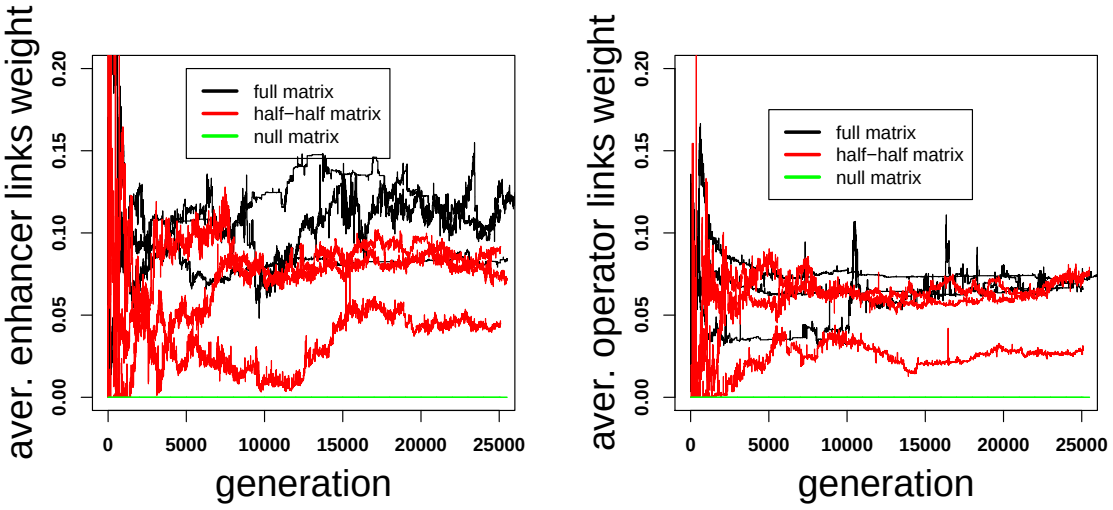
(a) number of enhancer links for the best individual

(b) number of inhibitory links for the best individual

Figure 13: Number of activation and repression links in the regulatory network for the best individual at each generation for all the simulations.

The organization of the regulation network appears more clearly when one looks at the interactions between the signaling protein and the rest of the network (figures 17 and 18). Both histograms (either activation of inhibition) are clearly different from the random ones: for the FULL context, the signaling protein has a strong inhibitory influence over many nodes in the network (figure 17, bottom line) while having only a little activation influence. This shows that, in the FULL context, the evolution has mainly selected direct influence of the signaling protein over the network. This is consistent with the previous results: since the network is only very slightly evolvable, the organisms were not able to develop a system involving the internal dynamic of the network to regulate their phenotype. The only solution is to use the external signal directly in order to regulate the transcription (even though the fitness curves show that this regulation is not very efficient, figure11(a)).

On the contrary, in the HALF-HALF context, the signaling protein is only locally connected to the network (figure 18). Therefore, the genetic network must transmit its influence toward all the proteins whose transcription rate needs to be modified during the organism's lifespan.



(a) average weight of the enhancer links

(b) average weight of the inhibitory links

Figure 14: Average weight of activation and repression links in the regulatory network of the best individual at each generation for all the simulations.

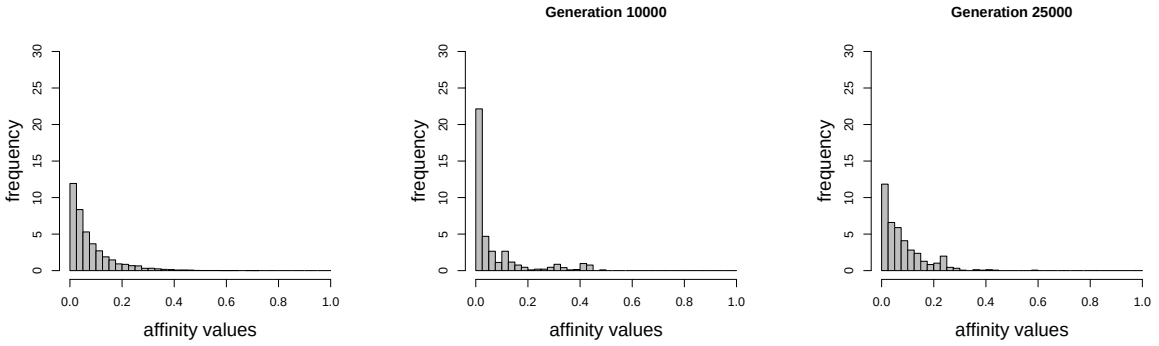
This is probably the reason why, in this context, the networks are composed of a larger number of enhancers than inhibitory links. Figure 11(a) shows that the result is indeed very efficient since HALF-HALF organisms have the smallest gap, hence the best fitness.

These results indicate that, in the FULL context, organisms have only developed a very simple (and almost inefficient) regulatory network. On the contrary, HALF-HALF organisms seem to develop a complex network. Nevertheless these histograms are not sufficient to understand the mechanisms of these complex networks, and so we will need to study their properties more precisely. To do so, we studied the final regulatory network of the best individual for the best simulation in order to see how it is structured. Results are presented in the next section.

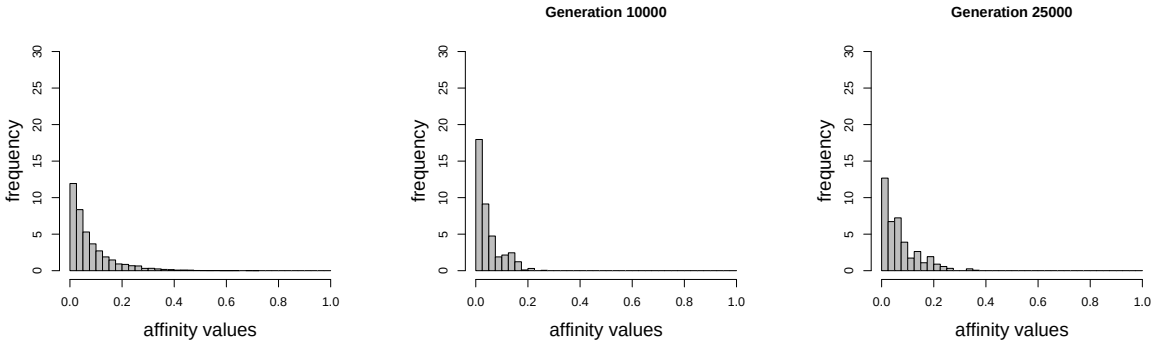
4.3 Analysis of a particular network

After 25000 generations, the HALF-HALF context presents a very efficient behavior: the gap value of the best individual is 0.0069 (whereas, without any regulation, the best possible gap is $\simeq 0.011$). It has a long genome ($\sim$ 10100 base pairs) with 51 genes (10 of them being transcription factors) and has developed a complex regulation network (figure 19).

Network dynamics have very good performance, as we can see in figure 20(a): a few time steps are enough to inhibit the subset of metabolic functions and to stabilize its behavior. In figure 20(b) we can see that after the arrival of the external signal, it only takes a few time steps to inhibit protein production and stabilize the network.



(a) Random distribution (b) Distribution at generation 10000 (c) Distribution at generation 25000

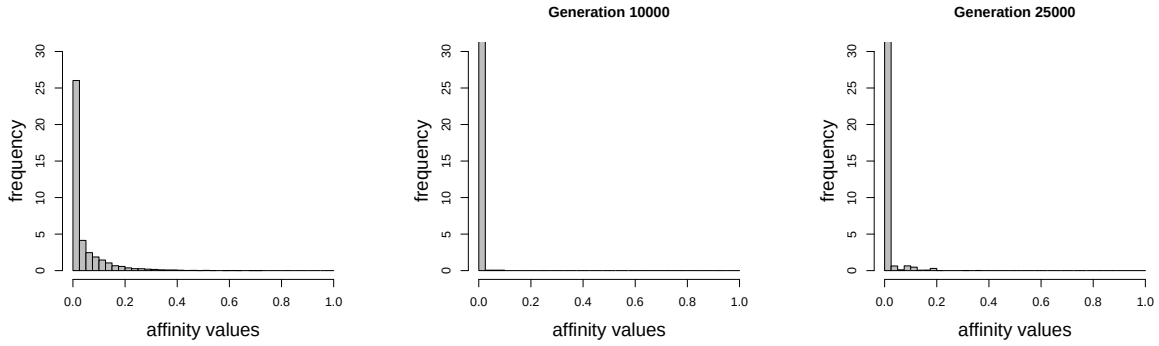


(d) Random distribution (e) Distribution at generation 10000 (f) Distribution at generation 25000

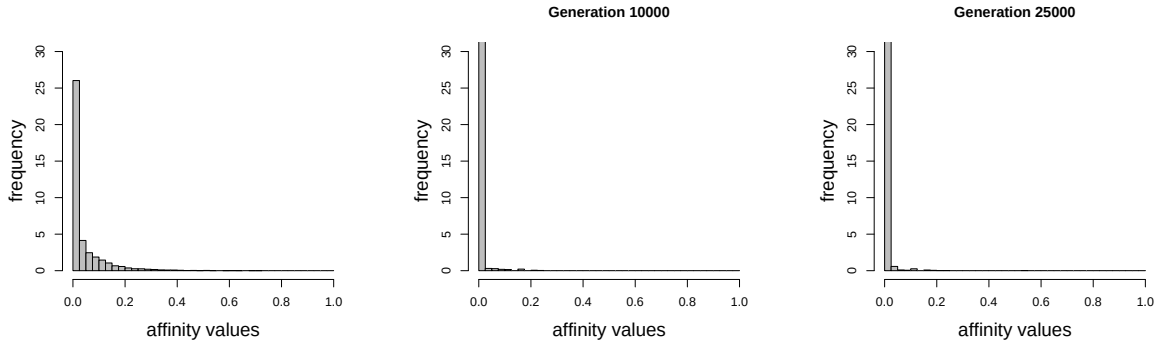
Figure 15: Distribution of interactions in the genetic networks in the FULL context (mean value for the three seeds). Top: Activators. Bottom: Inhibitory links. The first column presents the distribution obtained for random binding sites and random proteins of 21 amino-acids.

This network is highly connected: 47% of the links are active (791 active links vs 1688 possible links) – either positive (486 links, 56%) or negative (406 links, 44%)[8]. However, a large amount of these connections are still weak (data not shown) although some very active links have appeared in the network (mainly negative ones). In this experiment, the organisms have to adapt their metabolism when a signaling protein is introduced in the "cell". This protein can influence the transcription rate of genes either directly (by binding to one of its promoter's regulatory regions) or indirectly (by involving other intermediate regulators, i.e. transcription factors, in a complex regulation process). Indeed the regulation network does

[8]Note that the total number of links is not equal to the sum of enhancer links and inhibitory links. If a protein binds to both the operator and the enhancing region of a single promoter, we only count one regulatory link.



(a) Random distribution (b) Distribution at generation 10000 (c) Distribution at generation 25000

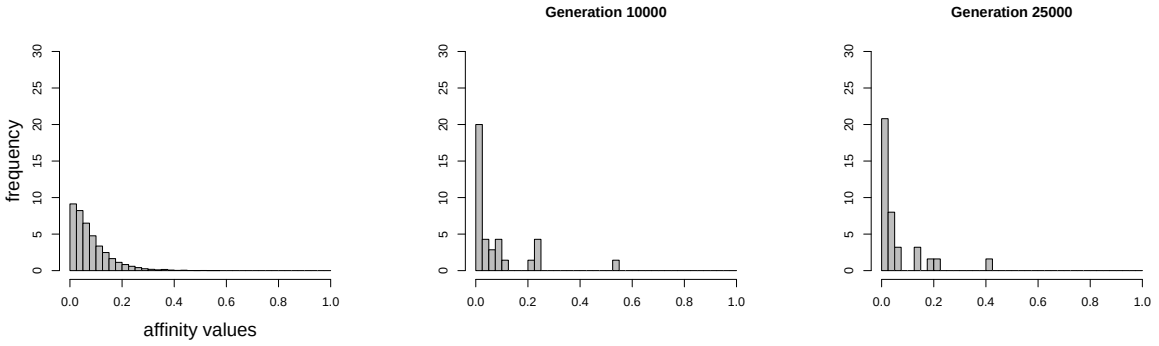


(d) Random distribution (e) Distribution at generation 10000 (f) Distribution at generation 25000

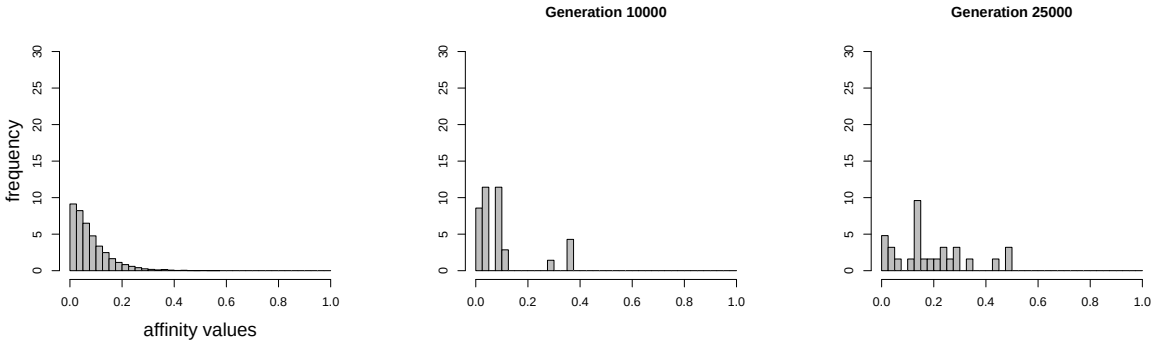
Figure 16: Distribution of interactions in the genetic networks in the HALF-HALF context (mean value for the three seeds). Top: Activation. Bottom: Inhibition. The first column presents the distribution obtained for random binding sites and random proteins of 21 amino-acids.

not need to be complex in order to be efficient.

In order to better understand the behavior of the regulation network, it is interesting to analyze the motifs that have emerged in the network [4, 26]. Table 2 shows the proportion of auto-regulation motifs in the evolved network. Clearly, the network has acquired more Positive Auto-Regulation (PAR) loops than Negative Auto-Regulation (NAR) ones. Yet, it has been demonstrated that Positive Auto Regulation slows down response time, decreases stability and increases variability [8]. Thus PAR can be positively selected. However the predominance of PAR may also be an indirect effect of the slightly higher proportion of enhancer links. Further analysis is therefore needed to distinguish these two hypotheses (selective hypothesis vs. neutral hypotheses) from each other.



(a) Random distribution (b) Distribution at generation 10000 (c) Distribution at generation 25000



(d) Random distribution (e) Distribution at generation 10000 (f) Distribution at generation 25000

Figure 17: Distribution of the influence of the signaling protein over the nodes of the genetic networks in the FULL context (mean value for the three seeds). Top: Activation. Bottom: Inhibition. The first column presents the distribution obtained for random binding sites.

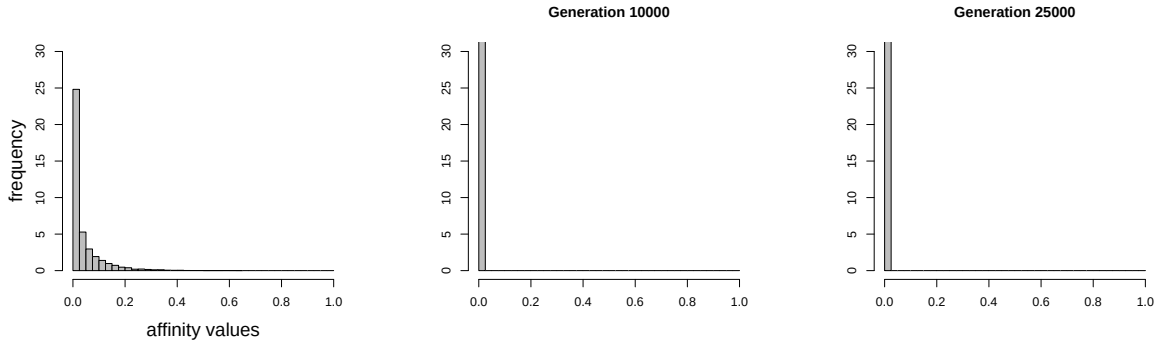
.

(PAR)	(NAR)	Isolated
8	3	40

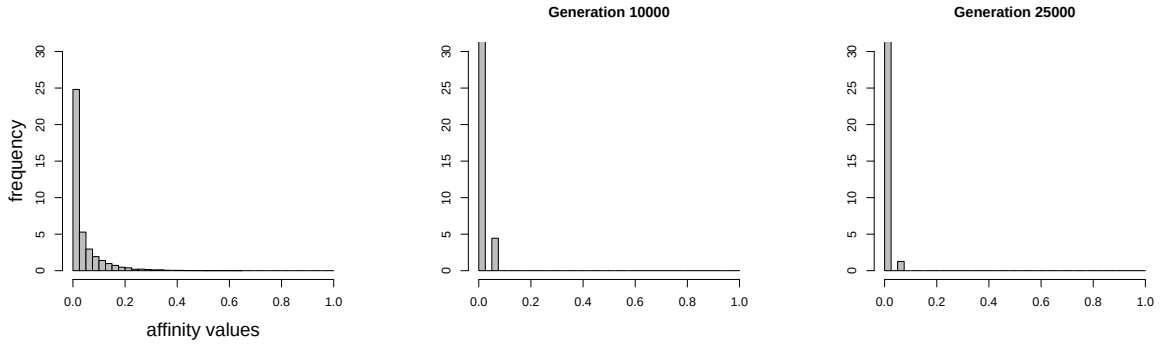
Table 2: Number of auto-regulation motifs in the network at generation 25000

Looking at two gene motifs (table 3), we can see the overrepresentation of Negative Feedback Loops. As discussed above for Auto-Regulation loops, this can be either a selective effect or a neutral effect. We now have to decipher between these two hypothesis.

Finally, when studying the regulatory network (figure 19), we have been surprised to find activation links from the signaling protein to a few nodes in the network. In fact



(a) Random distribution (b) Distribution at generation 10000 (c) Distribution at generation 25000



(d) Random distribution (e) Distribution at generation 10000 (f) Distribution at generation 25000

Figure 18: Distribution of the influence of the signaling protein over the nodes of the genetic networks in the HALF-HALF context (mean value for the three seeds). Top: Activation. Bottom: Inhibition. The first column presents the distribution obtained for random binding sites.

/	Activation	Inhibition	Positive Feedback Loop	Negative Feedback Loop	Positive Feedback Loop
620	303	227	38	64	23

Table 3: Number of binary motifs in the evolved network at generation 25000

when looking more precisely at the external protein links (Figure 21(a)) one can see that the signaling protein activates genes 1 and 22 and that protein 22 also inhibits gene 1. The whole

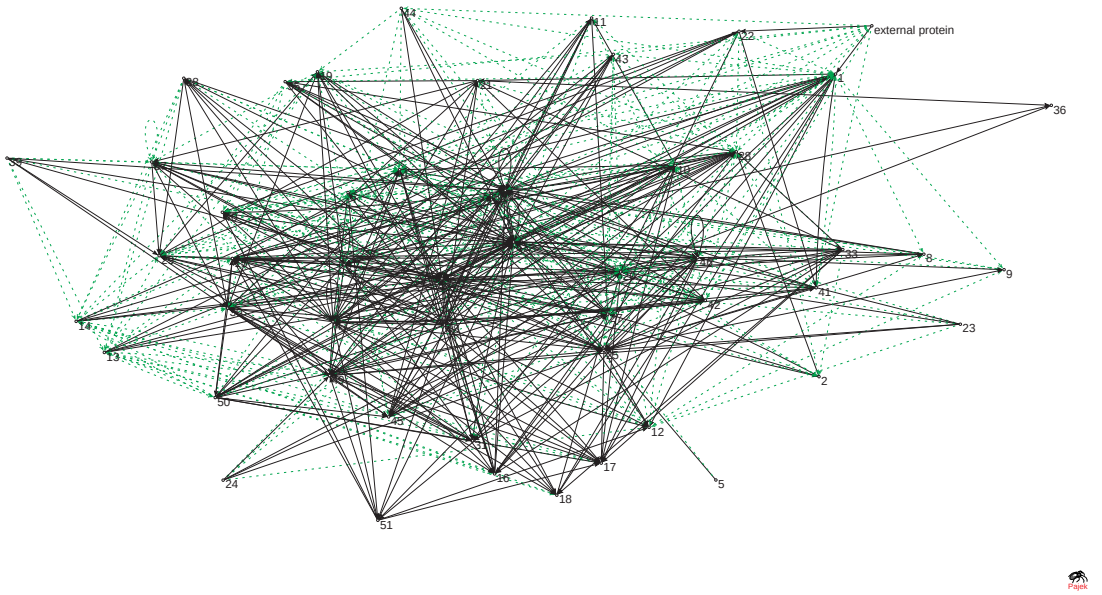
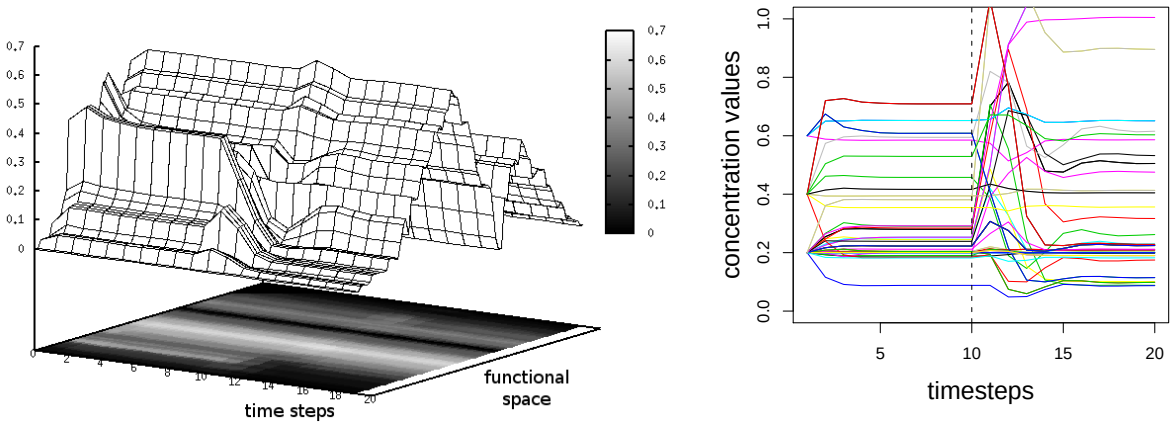


Figure 19: Regulatory Network of the best individual (HALF-HALF context). This image was generated with Pajek (Program for Large Network Analysis and visualization developed at Faculty of Social Sciences, Ljubljana). Inhibitory links are represented by dashed lines.

structure thus constitutes an incoherent Feed Forward Loop of type-1 (it is said incoherent because one side of the loop activates gene 1 while the other side inhibits it). This kind of loop has been well characterized in bacterial regulators (as found in *E. coli*) and it can carry out a response acceleration in dynamical functions [34].

Protein 22 can be considered as a transcription factor because it doesn't have any metabolic activity (the parameters of the protein are: $m = 0.28, w = 0.00$ and $h = -0.11$). The protein is weakly regulated (enhanced by protein 14 and the external signal; inhibited by protein 31). It also enhances the activation of 6 proteins and it inhibits five proteins, constituting a Single Input Modules (SIM) motif [4] (figure 21(a)). The architecture formed by the external protein, protein 22 and proteins in the SIM motif has been found by Cordero as the forerunner of FFL motifs avalanche [13].

Nodes 1 and 22 seem to act as sensory signal nodes. Protein 22, in particular, has mainly output links. It transmits the external signal (triggered by the external protein) to a subset of the other proteins. Protein 1 is a source of enhancing links but it only receives inhibitory influences. Even if this protein is enhanced by the external protein, it will be quickly repressed by incoming inhibitory links so its influence as an enhancer will be limited. This behavior can be seen in figure 21(b): when the external signal arrives to the cell, protein 1 is strongly enhanced but only a few steps later it is repressed, reaching to a steady state with a slightly higher concentration than before the arrival of the external signal.



(a) Variation of the phenotype during the 20 time steps of the individual's life

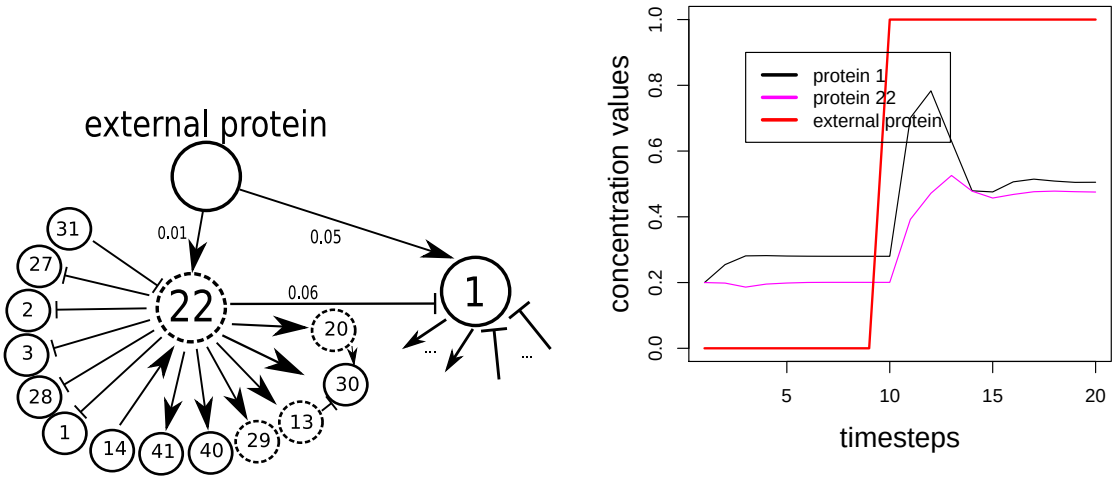
(b) Variations of the 51 proteins concentrations during the individual's life (the dashed line corresponds to the arrival of the external signal

Figure 20: Kinetic behavior of regulatory network. Best individual at generation 25000

These first experiments show that the RAevol model is able to produce "viable" regulation networks. However they also show that the high connectivity of the evolved regulation networks makes them very difficult to analyze. Hence, and as we are now aware of the influence of the affinity matrix on the connectivity of the resulting network, we can use these parameters to obtain more sparsely connected networks.

The evolutionary design of regulatory networks opens a lot of experimental directions. We are highly interested in investigating the mutational robustness of the networks: in RAevol, the mutational process is biologically realistic (i.e., mutations act at the genomic level rather than the regulatory level). Therefore, this model is particularly appropriate to better understand the complex relationship between the robustness of the organisms and the structure of their regulation networks.

Alternatively, we plan to extend the model by introducing stochasticity in the transcription process as well as stochasticity in the environment. We are particularly interested in the topology of the regulatory network: since the presence of Negative Auto-Regulation can reduce cell-cell variations, it can prevent the regulation network transmitting the stochasticity from the transcription process to the organism phenotype. On the contrary, variability can be enhanced by the creation of Positive Auto-Regulation (PAR) motifs [50]. Thus, we expect the number of PAR/NAR motifs to depend on the stochasticity of the environment: in highly stochastic environments, PAR should be positively selected for the phenotype to be stochastic too (bet-hedging). On the contrary, in a stable environment, NAR should be positively selected in order to reduce the effects of stochastic transcription.



(a) Schematic view of the subnetwork activated by the external protein. Nodes drawn with dashed lines are transcription factors

(b) Concentration of the external protein and of proteins 1 and 22. These three proteins all together create an Incoherent FFL motif [34]

Figure 21: Overview of the enhancing influences of external protein on the elements of the network

5 Open Issues and Future Work

RAevol is an integrated evolutionary model that provides experimenters with an insight to the complex adaptation mechanisms that prokaryotic organisms can adopt to face changing environments. It models the main features of the evolution and regulation of prokaryotes (although obviously in a very simplified way). In particular, it respects the different levels of organization of bacterial organisms as well as the interaction between those levels. When used to simulate the evolution of bacteria in a simple periodic and synchronized environment it proved to be a very valuable tool. Indeed the organisms not only developed simple strategies to regulate their metabolism but were also capable of building complex networks that allowed them to react efficiently to external events. However further investigation are needed to confirm these results. The behavior of the model itself also needs to be studied carefully: in this paper we showed that the mean connectivity of the regulation network of an individual has a strong influence on its evolutionary fate. We are now looking forward to conducting experiments with more sparsely connected networks (introducing more null values into the affinity matrix) in order to obtain biologically more plausible networks.

In the experiments presented here, the conditions our organisms had to face were quite simple. We now plan to test our model in more complex situations, in particular with non-synchronized environments where the optimal phenotype will vary in a more complex way. We suppose that, in such conditions, the organisms will develop more sophisticated regula-

tory networks, which would allow us to conduct a more in-depth study into the transcription motifs. In this paper we analyzed the first and second degree motifs but could go no further because of the high connectivity of the network. So, less connected networks would allow us to study the emergence of third degree motifs, FFLs, SIMs, ...

We will also analyze the topological characteristics of the networks: will they be scale-free [7]? will they adopt small-world [46] structure? Or will they have different characteristics depending on the specific features of the environment?

An open question is the relationship between the regulation network and the mutational robustness of the organisms. Does the regulation network enhance or reduce the organisms' robustness? Our firsts results suggest the latter but it clearly needs more investigation. A possible experiment would be to compare regulation networks evolved under different mutational constraints (e.g., different mutation rates).

Regarding the development of the RAevol model, our next step will be to introduce stochasticity in the transcription process and in the environment. From the work of Kussel and Leibler [33], we know that phenotypic noise may be selected in variable environments. Yet, it is an open question whether the regulation network will adopt a different structure depending on the necessity to amplify (or to reduce) the intrinsic transcription noise in order to adapt the phenotypic noise to the environmental conditions.

Acknowledgments: Antoine Coulon and Carole Knibbe for fruitful discussion on the model. Authors sincerely wish to thank the reviewer and Michael Parsons for their useful remarks and corrections. This work is supported by the Rhône-Alpes Complex Systems Institute (IXXI) and by the Spanish Ministry of Education (project number TIN2007-67148).

References

[1] C. Adami. *Introduction to artificial life*, (1998). Springer-Verlag New York, Inc., New York, NY, USA.

[2] C. Adami. *Digital genetics: Unraveling the genetic basis of evolution.* Nature Reviews Genetic, 7 (2006), 109–118.

[3] M. Aldana, E. Balleza, S. Kauffman, O. Resendiz. *Robustness and evolvability in genetic regulatory networks.* J Theor. Biol., 245 (2006), No. 3, 433-448.

[4] U. Alon. *Network motifs: theory and experimental approaches.* Nature Reviews Genetics, 8 (2007), No. 6, 450–461.

[5] W. Banzhaf. *Artificial regulatory networks and genetic programming.* In Rick L. Riolo, Bill Worzel, editors, Genetic Programming Theory and Practice, chapter 4 (2003), 43–62.

[6] W. Banzhaf. *On evolutionary design, embodiment and artificial regulatory networks.* Embodied Artificial Intelligence, (2004), 284–292.

[7] A. Barabasi , Zoltan N. Oltvai. *Network biology: Understanding the cell's functional organization.* Nature Reviews Genetics, 5 (2004), 101–113.

[8] A. Becskei , L. Serrano. *Engineering stability in gene networks by autoregulation.* Nature, 405 (2000), 590–593.

[9] T. Blible , L. Thiele. *A comparison of selection schemes used in evolutionary algorithms.* Evol. Comp., 4 (1996), 361–394.

[10] D. S. Burke, K.A. de Jong, J.J. Frefenstette, C.L. Ramsey, A. Wu. *Putting more genetics into genetic algorithms.* Evol. Comp., 6 (1998), 387–410.

[11] L. Cai, N. Friedman, X.S. Xie. *Stochastic protein expression in individual cells at the single molecule level.* Nature, 440 (2006), 358–362.

[12] S. Ciliberti, O.C.Martin, A. Wagner. *Robustness can evolve gradually in complex regulatory gene networks with varying topology.* PLoS Computational Biology, 3 (2007), No. 2.

[13] Otto X. Cordero , Paulein Hogeweg. *Feed-forward loop circuits as a side effect of genome evolution.* Molecular Biology and Evolution, 23 (2006), No. 10, 1931–1936.

[14] E. Dekel , U. Alon. *Optimality and evolutionary tunings of the expression level of a protein.* Nature, 436 (2005), No.7050, 588–522.

[15] P. Dittrich, J. Ziegler, W. Bazhaf. *Artificial chemistries: a review.* Artificial Life, 7 (2001), 225–275.

[16] S.F. Elena, R.E. Lenski. *Microbial genetics: Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation.* Nature Reviews Genetics, 4 (2003), 457–469.

[17] M.B. Elowitz, A.J. Levine, E.D. Siggia, P.S. Swain. *Stochastic gene expression in a single cell.* Science, 297 (2002), 1183–1186.

[18] P. Francois , V. Hakim. *Design of genetic networks with specified functions by evolution in silico.* PNAS, 101 (2004), No. 2, 580–585.

[19] J. Peter Gogarten, W. Ford Doolittle, Jeffrey G. Lawrence. *Prokaryotic evolution in light of gene transfert.* Molecular Biology and Evolution, 19 (2002), No. 12, 2226–2238.

[20] J. Hallinan , J. Willes. *Evolving genetic regulatory networks using an artificial genome.* 2nd Asia-Pacific Bioinformatics Conference, 29 (2004).

[21] A. Hintze , C. Adami. *Evolution of complex modular biological networks.* PLoS Computational Biology, 4(2008), No. 2.

[22] M. J. A. Van Hoek. *Evolutionary Dynamics of Metabolic Adaptation.* PhD thesis (2008), University of Utrecht.

[23] M. J. A. Van Hoek , P. Hogeweg. *The role of mutational dynamics in genome shrinkage.* Molecular Biology and Evolution, 24 (2007), 2485–2494.

[24] F. Jacob , J. Monod. *Genetic regulatory mechanisms in the synthesis of proteins.* J. Mol. Biol., (1961), No. 3, 318–356.

[25] N. Kashtan , U. Alon. *Spontaneous evolution of modularity and network motifs.* PNAS, 102 (2005), No. 39, 13773–13778.

[26] N. Kashtan, S. Itzskovitx, R. Milo, U. Alon. *Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs.* Oxford Bioinformatics, 20 (2004), No. 11, 1746–1758.

[27] C. Knibbe. *Structuration des génomes par sélection indirecte de la variabilité mutationnelle, une approche de modélisation et de simulation.* PhD thesis (2006), INSA - Lyon.

[28] C. Knibbe, G. Beslon, V. Lefort, F. Chaudier, J.-M. Fayard. *Self-adaptation of genome size in artificial organisms.* In Proc. of the 8th European Conference, ECAL 2005, 3630 (2005), 423–432.

[29] C. Knibbe, A. Coulon, O. Mazet, J.M. Fayard, G. Beslon. *A long-term evolutionary pressure on the amount of noncoding dna.* Molecular Biology and Evolution, 24 (2007), No. 10, 2344–2353.

[30] C. Knibbe, O. Mazet, F. Chaudier, J.M. Fayard, G. Beslon. *Evolutionary coupling between the deleteriousness of gene mutations and the amount of non-coding sequences.* J. Theor. Biol., 244 (2007), No. 4, 621–630.

[31] V. Kunin , C. A. Ouzounis. *The balance of driving forces during genome evolution in prokaryotes.* Genome Res., 13 (2003), No. 7, 1589–1594.

[32] P. Dwight Kuo, A. Leier, W. Banzhaf. *Evolving dynamics in an artificial regulatory network model.* In Parallel Problem Solving from Nature - PPSN VIII, 8th Int. Conf., Birmingham, UK, Sept. 18-22, 2004, Proceedings of LNCS, 3242 (2004),571–580.

[33] E. Kussell , S. Leibler. *Phenotypic diversity, population growth, and information in fluctuating environments.* Science, 309 (2005), No. 5743, 2075–2078.

[34] S. Mangan, S. Itzkovitz, A. Zaslaver, U. Alon. *The incoherent feed-forward loops accelerates the response-time of the gal system of escherichia coli.* J. Mol. Biol., 356 (2006), 1073–1081.

[35] H. H. McAdams, B. Srinivasan, A. P. Arkin. *The evolution of genetic regulatory systems in bacteria.* Nature Review Genetics, 5 (2004), No. 3, 169–178.

[36] P. Mendes, W. Sha, K. Ye. *Artificial gene networks for objective comparison of analysis algorithms.* Bioinformatics, 19 (2003), No. 2.

[37] C. Ofria , C. O. Wilke. *Avida: a software platform for research in computational evolutionary biology.* Artif. Life, 10 (2004), No. 2, 191–229.

[38] B. O'Neill. *Digital evolution.* PLoS Biology, 1 (2003), No. 1, 11–14.

[39] T.S. Ray. *An approach to the synthesis of life.* In Artificial Life II, volume XI (1991), 371–408.

[40] Y. Sanchez-Dehesa, L. Cerf, J. M. Pena, J.F. Boulicaut, G. Beslon. *Artificial Regulatory Networks Evolution.* In Proceedings of MLSB (2007).

[41] Y. Sanchez-Dehesa, J. M. Pena, G. Beslon. *RAevol, un modèle de génétique digitale des réseaux de régulation.* In Réseaux d'Interactions : Analyse, Modélisation, Simulation (RIAMS)(2007).

[42] O.S. Soyer , S. Bonhoeffer. *Evolution of complexity in signaling pathways.* PNAS, 103 (2006), No. 44, 16337-16642.

[43] K. Struhl. *Fundamentally different logic of gene regulation in eukaryotes and prokaryotes.* Cell, 9 (1999), No. 1, 1–4.

[44] F. Taddei, M. Radman, J. Maynard-Smith, B. Toupance, Pierre-Henry Gouyon, Bernard Godelle. *Role of mutator alleles in adaptive evolution.* Nature, 387 (1997), 700–702.

[45] S. A. Teichmann , M. M. Babu. *Gene regulatory network growth by duplication.* Nature Review Genetics, 36 (2004), No. 5, 492–496.

[46] D. J. Watts , S. H. Strogatz. *Collective dynamics of 'small-world' networks.* Nature, 393 (1998), No. 6684, 440–442.

[47] C. O. Wilke, J. L. Wang, C. Ofria, R. E. Lenski, C. Adami. *Evolution of digital organisms at high mutation rates leads to the survival of the flattest.* Nature, 412 (2001), 331–333.

[48] Joanne M. Willey, Linda M. Sherwood, Christopher J. Woolverton. *Microbiology.* Mc Graw Hill (2008), 7th edition.

[49] E. Yeger-Lotem, S. Sattath, N. Kashtan, S. Itzkovitz, R. Milo, R. Y. Pinter, U. Alon, H. Margalit. *Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction.* PNAS, 101 (2004), No. 16, 5934–5939.

[50] Y. T. Maeda Yusuke , M. Sano. *Regulatory dynamics of synthetic gene networks with positive feedback.* J. Mol. Biol., 359 (2006), 1107–1124.

[51] L. Zadeh. *Fuzzy sets as the basis for the theory of possibility.* Fuzzy sets and Systemps, 1 (1978), 3–28.