Numéro d'ordre 2011-0140
Année 2011

Sélection Indirecte en Évolution Darwinienne : Mécanismes et Implications

Thèse présentée par
David Parsons, Ingénieur INSA en Informatique

Devant
L'Institut National des Sciences Appliquées de Lyon

Pour obtenir
Le grade de Docteur

Formation doctorale
Informatique et Mathématiques

Soutenue le 8 décembre 2011 devant le jury composé de :

Wolfgang Banzhaf	Professeur, Memorial University of Newfoundland, Rapporteur
Guillaume Beslon	Professeur, INSA de Lyon, Directeur de thèse
Nicolas Bredèche	Maître de Conférences/HDR, Université Paris-Sud XI, Rapporteur
Alessandra Carbone	Professeur, Université Pierre et Marie Curie, Examinateur
Carole Knibbe	Maître de Conférences, Université Claude Bernard, Directrice de thèse
Marie-France Sagot	Directeur de Recherche, INRIA, Examinateur
Dominique Schneider	Professeur, Université Joseph Fourier, Examinateur

Numéro d'ordre 2011-0140
Année 2011

Indirect Selection in Darwinian Evolution: Mechanisms and Implications

Thèse présentée par
David Parsons, Ingénieur INSA en Informatique

Devant
L'Institut National des Sciences Appliquées de Lyon

Pour obtenir
Le grade de Docteur

Formation doctorale
Informatique et Mathématiques

Soutenue le 8 décembre 2011 devant le jury composé de :

Wolfgang Banzhaf	Professeur, Memorial University of Newfoundland, Rapporteur
Guillaume Beslon	Professeur, INSA de Lyon, Directeur de thèse
Nicolas Bredèche	Maître de Conférences/HDR, Université Paris-Sud XI, Rapporteur
Alessandra Carbone	Professeur, Université Pierre et Marie Curie, Examinateur
Carole Knibbe	Maître de Conférences, Université Claude Bernard, Directrice de thèse
Marie-France Sagot	Directeur de Recherche, INRIA, Examinateur
Dominique Schneider	Professeur, Université Joseph Fourier, Examinateur

Remerciements

Before I switch to french, I would like to warmly thank Wolfgang Banzhaf for accepting to review my work and to come accross all the way from Canada for my defence.

Nombreuses sont les personnes que je souhaite remercier au moment où j'achève ce travail. Un grand merci, tout d'abord, à Nicolas Bredèche pour avoir lui aussi accepté d'être mon rapporteur ainsi que pour ses critiques constructives de ce manuscrit.

Je souhaite également adresser tous mes remerciements à Alessandra Carbone, Marie-France Sagot et Dominique Schneider qui ont accepté de participer à mon jury.

Je tiens à exprimer ma gratitude à l'égard de tous mes anciens enseignants qui par leurs conseils et bien sûr leurs enseignements ont largement contribué à m'amener jusqu'ici, à Caroline, Mabrouka, Catherine, Caroline et Jeanine, toujours présentes dans les bons comme les mauvais moments et dont la bonne humeur et l'amabilité font légende.

Je remercie tous les membres de l'équipe Turing devenue Combining sans qui mon travail n'aurait sans doute pas été aussi agréable et motivant, avec une pensée particulière pour Yolanda, nos franches rigolades et nos concours de cuisine, pour Gaël avec qui j'ai tant partagé au cours de ces trois années et enfin pour Jules, Stephan et Bérénice qui m'ont fait l'amitié de relire ce manuscrit.

Un énorme merci à mes directeurs de thèse, Carole et Guillaume, pour avoir été extrêmement présents et disponibles durant ces trois années, pour m'avoir soutenu dans mes moments de doute, pour avoir su me montrer la bonne direction quand je m'égarais et surtout pour m'avoir donné le goût de la recherche en me proposant de travailler sur ce sujet fascinant.

Je pense bien sûr à mes parents et grands-parents qui m'ont toujours soutenu à tous égards et à Audrey, ma future femme qui m'a supporté notamment durant ces derniers mois et qui, par son soutien, m'a permis d'aller jusqu'au bout.

Enfin je remercie tous ceux qui m'ont soutenu et qui, si je n'ai pas pu les citer ici, se reconnaitront dans ces lignes. À vous tous, MERCI.

Abstract

In silico experimental evolution is a relatively new field of research that aims at studying the dynamics of evolutionary processes. Evolution experiments are conducted by simulating the evolution of populations of digital organisms in controlled conditions. Then, comparative analyses of the different outcomes of evolution depending on a reduced set of parameters allow the practitioner to shed light on intricate effects that would have been very difficult to identify otherwise.

The Aevol model is an *in silico* experimental evolution model that was specifically developped by Carole Knibbe to study the evolution of the structure of the genome. Using Aevol, a very strong second-order selective pressure towards a specific level of mutational variability of the phenotype was revealed: it was shown that since the survival of a lineage in the long term is conditioned to its ability to produce beneficial mutations while not losing those previously found, a specific trade-off between replication fidelity and evolvability is indirectly selected. The balance between replication fidelity and evolvability depends on the per base rate of local mutations, but also on the per base rate of chromosomal rearrangements and the genome size. Thus, a consequence of this indirect selective pressure is the central role played by the spontaneous rate of chromosomal rearrangements in determining the structure of the genome. More specifically, it was shown that because some rearrangements (large duplications and large deletions) have an impact not only around their breakpoints but on the whole sequence between them, non-coding sequences are actually mutagenic for the coding sequences they surround. The consequence is a clear trend for organisms having evolved under high rearrangement rates to have very short genomes with hardly any non-coding sequences while organisms evolving in the context of low rearrangement rates have huge, mostly non-coding genomes. Still, many questions remained open regarding both the precise conditions under which this indirect pressure can be involved, and its putative impact on other levels of organization such as the transcriptome and the proteome. Besides, in the experiments conducted by C. Knibbe, the breakpoints for rearrangements were chosen at random. Given the central role of chromosomal rearrangements in this process, a finer modelling of these rearrangements was badly needed to account for specificities of these events and in particular their sensitivity to sequence similarity.

Here, we modified the Aevol model to introduce an explicit regulation of gene expression as well as a sensitivity to sequence similarity in DNA recombination events. We observed that the effects of the second-order pressure mentioned above are very robust to modelling choices: they are similarly observed when gene regulation is made available, when rearrangements occur preferentially between similar sequences and even when a bi-

ologically plausible process of horizontal transfer is allowed. Moreover, the effects of this second-order selective pressure are not limited to the genomic level: high rearrangement rates usually lead to genomes that have many polycistronic RNAs, almost no non-coding RNAs and very simple regulation networks. On the contrary, at low rearrangement rates organisms have most of their genes transcribed on monocistronic RNAs, they own a huge number of non-coding RNAs and present very complex and intricate regulation networks.

These effects at different levels of organization can account for many features found on real organisms. Thus, the indirect selective pressure that was identified thanks to the Aevol model makes it possible to reproduce a large panel of known biological properties simply by changing the spontaneous rearrangement rate, making this pressure a good candidate for explaining these observations on real organisms.

List of personal publications

Published articles

- **D. P. Parsons**, C. Knibbe et G. Beslon. Homologous and Illegitimate Rearrangements: Interactions and Effects on Evolvability. *Proceedings of the Eleventh European Conference on the Synthesis and Simulation of Living Systems (ECAL 11)*, Paris, France. 2011.
- **D. P. Parsons**, C. Knibbe et G. Beslon. Aevol : un modèle individu-centré pour l'étude de la structuration des génomes. *Proc. MajecSTIC*, Bordeaux, 2010.
- G. Beslon, **D. P. Parsons**, Y. Sanchez-Dehesa, J. M. Pena et C. Knibbe. Scaling Laws in Bacterial Genomes: A Side-Effect of Selection of Mutational Robustness. *BioSystems* 102(1):32-40. 2010.
- G. Beslon, **D. P. Parsons**, J. M. Pena, C. Rigotti et Y. Sanchez-Dehesa. From Digital Genetics to Knowledge Discovery: Perspectives in Genetic Network Understanding. *Intelligent Data Analysis Journal* 14(2):173-191, 2010.
- **D. P. Parsons**, C. Knibbe et G. Beslon. Importance of the rearrangement rates on the organization of transcription. *Proceedings of the Twelfth International Conference on the Synthesis and Simulation of Living Systems (ALife XII)*, Odense, Denmark. 2010.
- Y. Sanchez-Dehesa, **D. P. Parsons**, J. M. Pena et G. Beslon. Modelling Evolution of Regulatory Networks in Artificial Bacteria. *Mathematical Modelling of Natural Phenomena* 3(2):27-66. 2008.

Conference abstracts

- C. Knibbe, **D. P. Parsons** et G. Beslon. Parsimonious Modeling of Scaling Laws in Genomes and Transcriptomes. *Proc. European Conference on Artificial Life, ECAL.* 2011.
- **D. P. Parsons**, C. Knibbe et G. Beslon. Influence of the rearrangement rates on the organization of genome transcription. *Integrative Post-Genomics*, Lyon, France. 2010.
- **D. P. Parsons**, C. Knibbe et G. Beslon. Influence of the rearrangement rates on the organization of genome transcription. *JOBIM*, Montpellier, France.2010.

- **D. P. Parsons**, G. Beslon, C. Knibbe, Y. Sanchez-Dehesa, J. M. Pena. Evolution of scaling laws in artificial regulation networks. *Integrative Post-Genomics*, Lyon, France. 2009.
- G. Beslon, Y. Sanchez-Dehesa, **D. P. Parsons**, J.M. Pena, C. Knibbe. Scaling Laws in Digital Organisms. *Proc. Information Processing in Cells and Tissues IPCAT'09*, Ascona, Switzerland. pp. 111-114. 2009.

Contents

List of Figures

List of Tables

Introduction

Since Charles Darwin's *Origin of Species* (Darwin, 1859) 150 years ago, it has been known that evolution takes place whenever three basic ingredients are put together: inheritance, variation and selection. Since then, many advances have been made in understanding the process of evolution and the underlying principles.
Today, the support of inheritance and the mechanisms underlying its expression and transmission are very well known: any living thing on earth owns a genome, usually encoded on a DNA (DeoxyriboNucleic Acid) sequence, which contains a certain number of genes, each of which will be translated into a protein through a transcription-translation process. This DNA sequence is replicated and transmitted to the next generation that is thus able to produce the same proteins, at least in the same environmental conditions. However, this genetic sequence is not totally stable, it is subject to physical and chemical stress that can damage it, thus causing modifications in the sequence, either locally (point mutations, small insertions or deletions) or at a genome-wide level, leading to chromosomal rearrangements. Rearrangements can reshape the whole chromosome, they can occur when the DNA is damaged to the point that the chromosome is cut into pieces. Some error-repair mechanisms allow the DNA segments to be reassembled but can fail to restore their original order, thus producing a rearranged chromosome. Rearrangements can also occur during DNA replication when the DNA-Polymerases make errors, or as a side-effect of the activity of transposable elements.
In comparison, the selection process could be thought of as trivial: a "simple" competition allowing for sorting between the products of genetic variability, the "fittest" organisms having a greater probability of reproduction than less fit ones. However, this selection process is a lot more intricate. Of course, if one considers two very different individuals in a population, one might have a direct physical advantage over the other: a blind eagle might find it more difficult to feed or breed than a sharp-sighted one. On the opposite, if one considers two phenotypically equivalent individuals, one could postulate that selection is blind-folded in that case and that only random genetic drift might be in action. This is not true in the long term: let us consider three phenotypically identical individuals that only differ through their tendency to produce genetic variations, for example because of their mutation rate. Mutations being mostly deleterious, an organism with a very high mutation rate will produce very few (if any) viable offspring. In extreme cases (above the "error threshold" – Eigen, 1971; Biebricher and Eigen, 2005), the lineage of such an organism undergoes more mutations than it can sustain and will quickly be driven to error catastrophe (Orgel, 1963) and extinction. As it is, it would be tempting to suggest that the lower the mutation rate of an organism, the more viable offspring it has and hence the higher its probability of "winning" the evolutionary competition. Yet, every now and then, a beneficial mutation can occur, resulting in a fitter offspring that will progressively

overcome the entire population. As a matter of fact, a beneficial mutation is more likely to occur in a moderately varying lineage than in a highly conservative one. Thus, there seems to be an optimal trade-off between the mutational robustness of the phenotype of an individual and its ability to evolve, namely its evolvability (Kirschner and Gerhart, 1998). Moreover, robustness and evolvability can themselves be selectionable traits. In other words, they can themselves be subject to evolution (Earl and Deem, 2004), leading to what is sometimes called "second-order selection".

As we have previously stated, genetic variation is usually the result of errors (mutations) in the DNA sequence. However, many error-correction mechanisms have evolved that can reduce by orders of magnitude the rate at which these errors occur. Yet, when looking at different kinds of organisms among the living kingdom, one can observe great differences in their spontaneous mutation rates: for example the spontaneous mutation rate of bacteriophage T4 was estimated to 2×10^{-8} per bp per replication (Drake, 1991) while that of *drosophila melanogaster* was estimated to 3.4×10^{-10} per bp per replication (Drake et al., 1998). This suggests that different life forms did not evolve error-repair systems to the same extent, which in turn raises several questions: why have some organisms evolved very effective error-correction mechanisms while others have hardly evolved any? Are these differences the result of mere chance or did a selective pressure drive the evolutionary process one way or the other depending on environmental conditions? While these questions are very difficult to answer, one can nonetheless imagine evolutionary explanations for these differences. One could argue, for instance, that the high mutation rates undergone by viruses allow them to evolve faster. Then, despite the immediate cost of having many non-viable offspring (and hence a lower fitness), in the long-term, lineages having higher mutation rates could evade the immune response of their hosts by evolving faster than the immune response itself. This is an example of indirect (or second-order) selective pressure towards evolvability. An example of second-order pressure towards robustness can be found in the structure of the genome itself when several genes overlap, sharing parts of their sequences, thus yielding a reduced mutational target. Direct forces that drive evolution (selection of the fittest) are thus coupled with more subtle, indirect (second-order) pressures toward *e.g.* robustness or evolvability.

The effects of such second-order selective pressures are very difficult to study, either *in vivo* or *in vitro*. Indeed, evolution is an intricate process that takes place on a very long time scale. It involves many different mechanisms that are entangled and whose effects can span several levels of organization and be observed differently on different time scales. If experimental evolution using real organisms is possible – the most famous example being R. Lenski's experiment on *Escherichia coli* that has been running for over 50.000 generations (Woods et al., 2011) – it remains extremely long and costly, which prevents it from being used extensively. Moreover, both real organisms and the environments they evolve in are very complex and it is nearly impossible to isolate the observed phenomena and hence to identify links of causality between specific environmental features and the product of evolution. Besides, the random nature of evolution makes it difficult, if not impossible, to reproduce events of interest that occurred by chance.

Comparative genomics approaches are a way to study evolution on long time scales. Genomes of different species are studied and compared with one another to reconstruct the evolutionary history of life as we know it today. Algorithms based upon the similarities between sequences from different species allow biologists to identify and characterize

mechanisms that take part in the evolutionary process. However, these approaches are solely based on a snapshot of contemporaneous sequences, upon which the effects of many different mechanisms and pressures are superimposed, making the puzzle very difficult to disentangle. Besides, fossil records being quite sparse, it is usually difficult to confirm the results obtained with these approaches. Finally, to quote John Maynard Smith, "We badly need a comparative biology. So far, we have been able to study only one evolving system and we cannot wait for interstellar flight to provide us with a second" (Maynard Smith, 1992). Indeed, as we have already emphasized, evolution is highly undeterministic, meaning that life could have been very different from life as we know it. To study evolution as a process, we need several repetitions of this process, which, as a matter of fact, we don't have.

Interestingly this particular quote from John Maynard Smith continues: "If we want to discover generalizations about evolving systems, we will have to look at artificial ones". In other words, as "real" biology provides us with only one single instance, the only way we can compare different instances is by creating artificial ones and hence by building models of evolution that we can simulate a great number of times.

Population genetics and the theory of quasispecies (Eigen et al., 1989) are a way to tackle the problem. By building mathematical models of evolution, one can conduct analytic studies of the effects of different parameters on the outcome of evolution. However, these models usually rely on very strong assumptions such as infinite population sizes or single gene "organisms".

An alternative approach to study evolution is to build an explicit model of evolution and run simulations with it. This approach is referred to as *digital genetics* (Adami, 2006). Terminologies of artificial or *in silico* evolution are also common. In the last 20 years, digital genetics have proved to be of major interest in unravelling universal mechanisms governing evolution, especially when it comes to second-order pressures towards robustness and / or evolvability (Wilke et al., 2001; Adami, 2006).

The Aevol model in particular was developed by Carole Knibbe during her PhD thesis (Knibbe, 2006) to study the evolution of the structure of the genome, making it the perfect candidate to observe indirect selection and study its consequences on evolution and genome organization. Using this model, Knibbe et al. (2007a) identified a strong second-order selective pressure towards a specific level of mutational variability of the phenotype. In these experiments, which are presented in more detail at the end of chapter I, the mutation rate was identified as a major determinant of the amount of non-coding DNA. A mathematical analysis of the mechanisms underlying genetic variation in the model revealed that this effect was due to the mutagenic effects of non-coding sequences for their neighbouring genes when chromosomal rearrangements are involved. In all the experiments presented in Knibbe et al. (2007a), the genome was shaped by evolution in such a way that the best individual in the population produced on average one neutral offspring, *i.e.* one offspring that shared the exact same phenotype as its parent's.

These seminal results shed light on a strong indirect pressure on the very structure of the genome. They are of major importance because i) it would have been very difficult to identify these pressures using another method than *in silico* evolution, and ii) they suggest hypotheses that are often overlooked when it comes to understanding particular structures in real genomes. These results showed that even when non-coding sequences have no direct impact whatsoever on the phenotype or fitness, the genome of evolved

organisms can still contain a huge proportion of them. Now if second-order pressures can lead to highly non-coding genomes single-handed, they may very well be able to produce other structures of interest or some kind of modularity upon the genome. Indirect selective pressures could also be proposed to have an impact on further levels of organization of the cell such as, for instance, the transcriptome or even the proteome. On the other hand, this raises the question of the robustness of these results. Even though many different *scenarii* were tested, showing that this pressure applies regardless of the shape or "complexity" of the environment, of the intensity of selection or of the degree of pleiotropy of proteins (*i.e.* the range of biological processes it contributes to), some questions remain: would the identified pressure still be involved in a system in which an additional degree of freedom is provided by a process of regulation of gene expression, and if so, what would be its impact on the structure and complexity of the regulatory network itself? Of utmost interest are the questions of homology driven rearrangements and of horizontal transfer: most chromosomal rearrangements occur between similar sequences. Yet, in Knibbe et al. (2007a), computational limits had made it impossible to account for sequence similarities in the chromosomal rearrangements model, so that rearrangement breakpoints were actually chosen at random along the chromosome. Finally, can horizontal transfer turn the observed effects off by providing a way to evade the problem of linkage disequilibrium (Sniegowski et al., 2000)? These questions are the starting point of this PhD thesis. To tackle them, we improved the implementation of the model to enable longer experiments (from a few tens of thousands of generations to up to millions of generations when needed) and conducted further experiments with both the original model and with two extensions of it: one including an explicit regulation process and the other, a more detailed model of chromosomal rearrangements based on sequence homology and allowing for a plausible horizontal transfer mechanism.

This manuscript is organized into five chapters. The first three chapters are dedicated to the model in its former version as well as its R-Aevol extension. In the first chapter, after a state-of-the-art of *in silico* experimental evolution, a detailed description of the model will be provided, followed by a presentation of its usage and the seminal results obtained with it. In chapter II, the causes and consequences of the second-order pressure identified by Knibbe et al. (2007a) will be discussed, focusing on the respective impact of mutations and rearrangements and on the effects of this pressure at the level of the transcriptome, showing in particular that operon structures can arise when they are not expected, as a vector towards genome streamlining. The third chapter will be dedicated to R-Aevol, an extension of the model developed by Yolanda Sanchez-Dehesa (Sanchez-Dehesa, 2009) in which an explicit gene regulation process was introduced to study the evolution of gene regulation networks. Experiments we conducted with R-Aevol showed a very strong effect of second-order selection on the size and complexity of regulation networks. Chapter IV describes another extension of the model, in which a sensitivity to sequence similarity was introduced in the chromosomal rearrangement process in order to investigate the role of both homologous and nonhomologous rearrangements in the evolution of genome structure. Finally, in chapter V, we will describe a model of homology-driven horizontal transfer and discuss both the effects of homologous and nonhomologous recombination on evolution, and the impact of transfer on the indirect pressure for a specific level of robustness and evolvability we have already mentioned.

This work shows the ubiquity of this second-order pressure, which acts as a strong de-

terminant not only of the structure of the genome, but also of that of the transcriptome and, at least to some extent, of the proteome. Moreover, this pressure seems to be very robust to modelling choices, its effects being consistently observed regardless of whether gene expression can be regulated or even of whether genetic material can be exchanged between lineages. As its effects span several levels of organisation, this pressure allows a large panel of biological structures observed in real organisms to be parsimoniously reproduced by acting on the rearrangement rate alone.

Chapter I

The Aevol Model

1 Introduction

The Aevol model is a *digital genetics* model (Adami, 2006) developed in the LIRIS lab to study the structuration of the genome in a Darwinian evolution process. Digital genetics models simulate, in a reasonable computational time, the evolution of a population of artificial organisms in a controlled environment. The typical use of these models is very similar to "wet" experimental evolution procedures (Elena and Lenski, 2003), so that it is also referred to as *in silico* experimental evolution: populations of organisms are initialized and left to evolve in controlled conditions. Then, by observing the products and the dynamics of the evolutionary process in different conditions and by comparing them, one can unravel the direct or indirect pressures that constrain the structure of the genome. However, because real organisms have a much greater generation time than artificial ones, time scales are fundamentally different. Thus, when a "wet" evolution experiment is very expensive and can last for decades (Blount et al., 2008), a simulated one takes only a few days or weeks and has a very reduced cost.
In this chapter, the state-of-the-art of *in silico* experimental evolution will be presented, followed by a focus on the Aevol model. A brief overview of Aevol will provide the reader with a general understanding of the whole process, then every stage of this process will be developed in more detail, focusing particularly on the specificities of Aevol with regard to other digital genetics models. Finally, the seminal results obtained with Aevol, the starting point of this PhD, will be presented.

2 *In Silico* Experimental Evolution: State of the Art

In silico experimental evolution is a relatively new field of research that emerged about 20 years ago. Populations of virtual organisms are placed in a virtual environment in which they compete to reproduce. Each organism owns some kind of genetic material which is interpreted by dedicated programs to compute its phenotype, which itself is the basis of

a selection process. Finally, this genetic material undergoes variation, usually during the reproduction process.
From a programmer's point of view, digital genetics models are very close to Genetic Algorithms. However, their respective goals differ greatly: on the one hand, genetic algorithms *use* our knowledge of the evolutionary process in order to find a solution to an engineering problem (in fact by *evolving* a solution), and on the other hand, the aim of *in silico* experimental evolution is rather to *study* the evolutionary process itself. A digital genetics model must thus put all the ingredients for Darwinian evolution together in a simple model, to allow the practitioner to study the emergence of particular properties, depending on different parameters.
This global scheme can be derived in many ways depending on the biological questions that are to be addressed. Different types of genetic material can be used, from sequences of instructions or nucleotides to pools of genes. The information contained by this genetic material can then be processed by different artificial chemistries (Dittrich et al., 2001) to fulfil a specific task such as resource- or data-processing. Selection can also take several forms, either synchronous or asynchronous, local or global, based directly on a fitness value or depending on the rank of the individuals in the population. Finally, the genetic operators that can be implemented strongly depend on how the genetic information is encoded: formalisms using sequences, and in particular sequences of nucleotides, can use very realistic genetic operators, while formalisms whose genetic material is more abstract are doomed to use highly specific operators as well.
I propose here a brief review of the different kinds of formalisms used in *in silico* evolution using a terminology inspired by that proposed by Hindré et al. (2011). Genetic algorithms used for engineering purposes are outside the scope of this review because their goal is not to study evolution but rather to use it for optimization purposes.

2.1 The "Program" Formalism

The first digital genetics models were "Program" models. In this class of models, the virtual organisms are actually programs that compete for computational time and possibly memory space in a virtual computer with an *ad-hoc* operating system. The genome of these individuals is the sequence of instructions that will be executed during their life time. Because the genome is a sequence, these models are well suited to study the evolution of some structural aspects of genomes such as gene-clustering. However, as the genome codes directly for a sequence of instructions without any notion of genes or proteins, it is difficult to compare them with that of real genomes.
Tierra (Ray, 1991) was the first model to be developed. In Tierra, organisms are not evaluated according to their performance of a predefined task, which makes it a model of real open-ended evolution. Organisms directly compete for survival by evolving better ways of reproducing. Hence, as in natural evolution, there is no need for a fitness value for the system to operate, fitness is simply a measure the experimentalist can use. In addition to the expected optimizations that were found by evolution, experiments using Tierra showed the spontaneous emergence of parasites and hyper-parasites as well as a certain kind of sociality (Ray, 1991, 1992). Tierra was also used as a test case for the classical successive substitution interpretation of evolution in a chemostat (Yedid and Bell, 2001), an interpretation that could be thought of as short-sighted.

Avida (Adami and Brown, 1994) is probably the most widely used model using the "Program" formalism. Contrary to Tierra, organisms are isolated from one another, thus being protected against parasitic attacks. They are given specific tasks (logical operations) to perform and are rewarded additional computational time accordingly. Thus, in Avida, evolution is no longer open-ended. Many works have used Avida mostly tackling the evolution of evolvability and robustness (Wilke et al., 2001; Wilke and Adami, 2003; Elena and Sanjuan, 2008), but also the evolution of complexity (Lenski et al., 2003) or modularity (Misevic et al., 2006) as well as adaptive radiation (Chow et al., 2004).
Musso and Feverati (2011) proposed a model of Turing Machine evolution. Using this model, they showed that the amount of active code that can be maintained by selection admits an upper bound – the error threshold (Eigen, 1971) – and that whatever the mutation rate imposed on the organisms, evolution pushes this amount of coding sequences toward the error threshold corresponding to this mutation rate.

2.2 The "MorphoElements" Formalism

In this formalism, the genome of an organism codes for body parts (morphological elements) interconnected by joints upon which forces can be applied either directly or through "muscle" elements. The behaviour of the organism is also encoded on the genome and co-evolve together with the morphological properties. Karl Sims' Creatures (Sims, 1994b) as well as Framsticks (Komosiński and Ulatowski, 1999) are probably the most widely known and used models of this kind. Both models were used to study the evolution of morphology, in particular when several organisms are placed in a situation of co-evolution, either through competition or predator-prey interactions (Sims, 1994a; de Back, 2006). This formalism is very much in vogue in the *Artificial Life* community and models of this kind are very numerous, including the GOLEM project (Pollack and Lipson, 2000), Blindbuilder (Devert et al., 2006) and Josh Bongard's robots (Bongard and Paul, 2001; Bongard, 2010), each of which has its particular interest. The main drawback of these systems is the lack of realism of the genomes, which are very specific and often recursive (in Creatures for instance, the genome takes the form of a recursive directed graph), which makes it very difficult to disentangle the effects of the evolutionary process itself from those caused by implementation choices. Moreover, as we have previously stated, such specific "genetic" encoding implies very strong constraints on the mutation operators that can be used, which have to be very specific and can hence hardly be compared with real mutation mechanisms. It must be mentionned that, in the last few years, several authors have focused on developmental approaches (Bongard and Pfeifer, 2003; Devert et al., 2007; Harding et al., 2010) that prove to be more evolvable than direct morphological encoding. However, this is outside the scope of the present work, that focuses on the evolution of the structure of the genome.

2.3 The "Network" Formalism

Models using the "Network" formalism have been developed to study specifically the evolution of gene regulation networks. These models are very distinctive in that the organisms are themselves, networks, represented in the form of connection weights or node functions. There is no explicit genome level, neither are there intermediate levels such as the tran-

scriptome or the proteome that, in real organisms, give rise to regulatory networks. It is indeed the interactions between cis- or trans- acting elements throughout these different levels of organization that engender regulation in real organisms. In models of the "Network" kind, the network is hard-wired and mutations directly modify its nodes' behaviour or connection weights. Again, these mutation operators being very specific, it is difficult to compare them to real mutation mechanisms. Nevertheless, important results were obtained using "Network" models, mostly regarding the evolution of network robustness and evolvability as well as modularity (Kashtan and Alon, 2005; Kashtan et al., 2007, 2009; Wagner, 1996; Espinosa-Soto and Wagner, 2010; Siegal and Bergman, 2002; Martin and Wagner, 2008; Ciliberti et al., 2007; Draghi and Wagner, 2009; Azevedo et al., 2006).

2.4 The "Allelic" Formalism

In the "Allelic" formalism, the genome is made up of a fixed number of loci, each of which is associated to a predefined set of possible alleles. Each allele of a gene is associated to a direct contribution to the organism's fitness that has to be predetermined arbitrarily. This formalism has been used to study the evolution of mutator alleles in different conditions and its effect on the increase of the average fitness of the population (Taddei et al., 1997; Tenaillon et al., 1999). Allelic models can also be used with an infinite number of alleles to explore the behaviour of a system under neutral drift, for example the conditions of divergence leading to speciation (Hanage et al., 2006). Many features of "Allelic" models are *ad-hoc* by essence, for instance many assumptions have to be made regarding the respective frequencies and impacts of beneficial, deleterious and lethal mutations. These models are also very abstract and do not capture features of any specific organism, rather, they account for the general dynamics of the evolutionary process and are thus best-suited for testing general processes such as the evolution of mutation rates.

2.5 The "String-of-Pearls" Formalism

In this formalism, which is also often referred to as the "Beads-on-a-String" formalism, the genome consists of a sequence of elements that usually represent genes but can also be of different natures, including transcription factor binding sites, retroposons and repeats. This formalism is flexible in terms of gene order and allows for consistent large chromosomal rearrangements. Important results were obtained using this kind of model, regarding the evolution of evolvability through modularity (Crombach and Hogeweg, 2007), the evolution of gene regulation networks in alternating environments (Crombach and Hogeweg, 2008), sympatric speciation (Tusscher and Hogeweg, 2009) and resource cycling in ecosystems (Crombach and Hogeweg, 2009).

2.6 The "Sequence of Nucleotides" Formalism

Here, the genome takes the form of a variable-length sequence of characters, each character representing a nucleotide. Specific predefined motifs are used as signal sequences (analogous to *e.g.* promoters) to detect the coding sequences upon the genome, the remainder being strictly non-coding. Because the structure of the genome is similar to that of real genomes, all the possible mechanisms of genetic variation found in real organisms

can be modelled in a realistic way, including chromosomal rearrangements and horizontal transfer. Furthermore, the explicit modelling of transcriptional and translational processes allow for the study of different levels of organization besides that of the genome, most notably the transcriptome and the proteome. Models of this kind have allowed to shed light onto the evolution of non-coding DNA and gene number (Knibbe et al., 2007a, 2008), the evolution of gene regulation networks (Kuo et al., 2006; Beslon et al., 2010b) and of metabolism (Flamm et al., 2010; Ullrich et al., 2011). It has also been proposed that such models could be used to generate benchmarks to test *e.g.* gene network inference strategies (Mattiussi and Floreano, 2007; Marbach et al., 2009; Beslon et al., 2010a).

2.7 Conclusion

Among these formalisms, none is systematically better than the rest. In fact, which formalism is best highly depends on the question one wants to address [1]. However, many questions require that genetic variation mechanisms be finely modelled, which can only be achieved through the "Sequence of Nucleotides" formalism. This is particularly true in our specific case: since we are interested in the evolution of the structure of the genome, we must model the genome in a very realistic way.

3 Overview of the Aevol Model

Aevol is a digital genetics model following the "Sequence of Nucleotides" formalism. It was developed by Guillaume Beslon and Carole Knibbe to study the evolution of the structure of the genome (Knibbe, 2006; Knibbe et al., 2007a,b). Because the goal is to use the model to study specifically the evolution of the structure of the genome, the model must be very realistic on that particular level. Not only must the genome itself have a biologically plausible structure, but also the genetic operators. Both local (point mutations and indels) and global (chromosomal rearrangements and horizontal transfers) genetic operators must be modelled, and both the prerequisite for the corresponding events and their effects must be highly similar to those of real genetic variation mechanisms. This biologically inspired genome must then be interpreted in terms of adaptation through a simple artificial chemistry (Dittrich et al., 2001) that can be simulated in a short computational time so that it can be integrated in an evolutionary loop. Interestingly, if the model has to be very realistic on the level of the genome, the further we get from this level, the more we can allow the model to be abstract. Indeed, as we are not interested in the phenotypic nor the ecological level, we do not need a very precise model at these levels. This will allow for a computationally tractable model of evolution featuring a very realistic genome level.
From a computer science point of view, Aevol is an individual-based evolutionary model that simulates the evolution of a population of N artificial organisms. At each generation, all the individuals are evaluated and compete for reproduction through a roulette wheel selection process. N offspring are produced by error-prone replication, forming the new population that will replace the former one. The initial generation is usually made up

1. This consideration is actually general to any modelling approach, the kind of model as well as its degree of precision or realism all depend on the question one wants to "ask" the model.

of N clones of a randomly generated individual with a basic prerequisite that it owns at least one "good" gene.
In Aevol, each organism possesses its own genome, whose structure is greatly inspired by bacterial genomes. It is organized as a circular double-strand binary string containing a variable number of genes most likely separated by non-coding sequences (figure I.1-a). Aevol being a "Sequence of Nucleotides" model, genes in Aevol are not defined by their position on the genome (locus specific). They are identified thanks to a set of predefined signalling sequences (figure I.1-b) and are translated into abstract proteins during an explicit transcription-translation process (figure I.1-c) using a predefined genetic code (figure I.1-d). Metabolic processes are defined within a metabolic function space Ω and each protein can either realize or inhibit a particular set of abstract biological processes with a certain possibility degree. Then, the metabolic activity of each protein is computed through three parameters m, w and h (figure I.1-e) characterizing a fuzzy set that represents the possibility degree at which each metabolic process is realized (figure I.1-f). This computation of the activity of proteins could be thought of as a kind of "folding" process. Once all the proteins have been characterized, their individual activities are combined, thus giving rise to the organism's phenotype (figure I.1-g) that is expressed in the same space as protein activities. Finally, this phenotype is compared to a predefined environmental target (also expressed in the same space) to determine how fit the organism is, *i.e.* to compute its fitness.

4 Proteins, Phenotypes and Environments: a simple Artificial Chemistry

To model the activity of proteins and the corresponding phenotype, we defined a simple artificial chemistry (Dittrich et al., 2001) that describes the metabolism of an organism in a mathematical formalism. We assume that there is an abstract space Ω of all the biological processes an organism could possibly accomplish. In the model, $\Omega = [0, 1]$ is a one-dimensional interval, so that a biological process is simply represented by a real number between 0 and 1. In this "metabolic space", the phenotype P of an organism is represented by a fuzzy set that expresses the efficacy with which this organism realizes each biological process in Ω (figure I.2). This efficacy is expressed in a fuzzy set formalism: the possibility theory. Thus, a protein's efficacy will also be referred to as the "degree of possibility" with which it realizes a particular process.
An organism is made up of proteins, and each protein is involved in a subset of biological processes, either contributing to their realization or inhibiting it with a possibility degree ranging between 0 and 1. The activity of a protein is then characterized by a function associating a certain possibility degree to each process in Ω. For reasons of simplicity, Aevol uses piecewise-linear functions with a symmetric triangular shape. Thence, only three numbers are needed to fully characterize the metabolic activity of a protein: the position m of the triangle on the biological process axis, its height h and its half-width w. A predefined genetic code (table I.1) associates each possible triplet of nucleotides (codons) with an abstract Amino-Acid (AA) that will have an impact on the value of one of the parameters m, w or h. The coding sequence of a gene hence consists of three interlaced binary sequences, each of which codes for one of the parameters of the protein

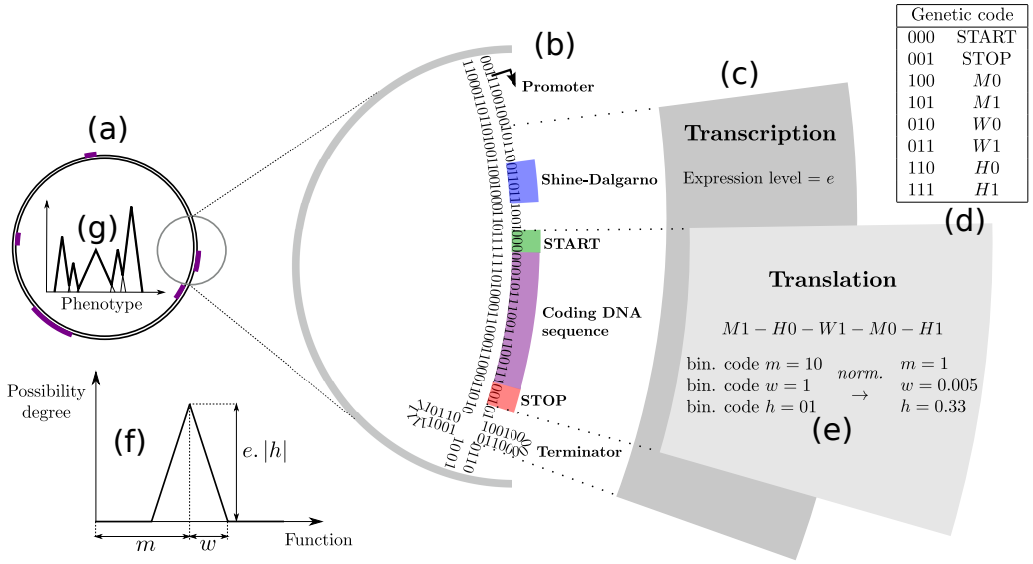


Figure I.1 – In Aevol, each organism possesses a circular double-strand binary genome **(a)** along which coding sequences are identified thanks to predefined signalling sequences **(b)**. Promoters and terminators mark the boundaries of transcribed sequences, *i.e.* RNAs **(c)** within which coding sequences can in turn be identified between a Shine-Dalgarno-Start signal and an in-frame Stop codon. The sequence thus identified will then be translated into the primary sequence of the corresponding protein thanks to a predefined genetic code **(d)**, this primary sequence being in turn interpreted as the m, w and h parameters of the protein **(e)**. Proteins, phenotypes and environments are represented similarly through fuzzy sets that associate a possibility degree to each possible metabolic function. For simplicity reasons, a protein's metabolic contribution takes the form of a piecewise-linear function with a triangular shape, the m, w and h parameters corresponding respectively to the position, half-width and height of the protein's metabolic activity **(f)**. All the proteins of the organism are then combined to compute the phenotype **(g)** that, once compared to the environment target, can be used to compute the fitness of the individual.

(see figure I.1-e). Finally, each of these sequences is interpreted as a real value using the Gray code and is then normalized within the domain of definition of the parameter. The thereby defined protein then contributes to the range $[m - w, m + w]$ of metabolic processes, either realizing or inhibiting them, with a preference for the processes closest to m (for which the highest efficacy h is reached). Whether the protein actually realizes the function or inhibits it depends on the value of the parameter h: a positive h will yield the realization of the functions and a negative value will inhibit them. In this framework, different types of proteins can co-exist, from highly efficient and highly specialized ones (small w, high h) to polyvalent but poorly efficient ones (large w, low h).

In the model, the environment is indirectly represented by a phenotypic target: the fuzzy set E, defined on Ω, that represents the optimal degree of possibility for each biological

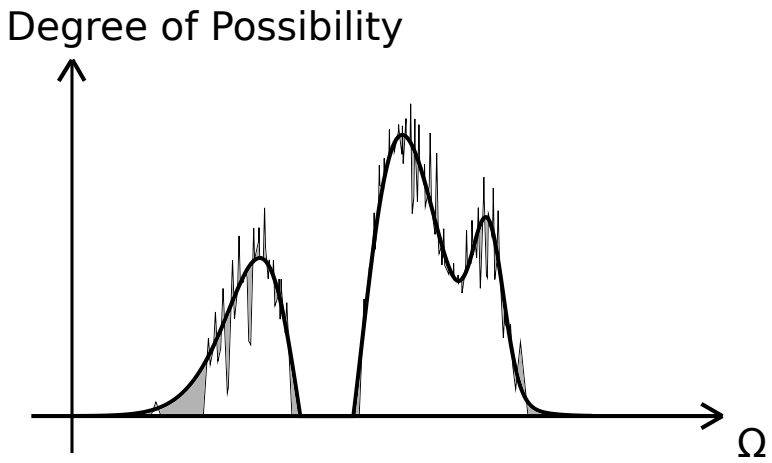


Figure I.2 – Fuzzy set representation of an organism's phenotype (thin line) and of the environmental target (thick curve). Each possible metabolic process in Ω is realized with a certain degree of possibility, that of the environmental target representing the optimal value. The grey area between the two curves represents the metabolic error (or gap) g of the organism. The smaller this metabolic error, the fitter the organism.

Codon	AA	Meaning
000	START	Marks the beginning of a gene (in conjunction with an upstream Shine-Dalgarno sequence)
001	STOP	Marks the end of a gene
100	M0	Adds a 0 bit to the binary sequence for parameter m
101	M1	Adds a 1 bit to the binary sequence for parameter m
010	W0	Adds a 0 bit to the binary sequence for parameter w
011	W1	Adds a 1 bit to the binary sequence for parameter w
110	H0	Adds a 0 bit to the binary sequence for parameter h
111	H1	Adds a 1 bit to the binary sequence for parameter h

Table I.1 – The genetic code that was used in all the experiments presented in this manuscript. Note that there is no redundancy.

process in this particular environment. To evaluate an individual, we compare its phenotype P to this environmental target E. The geometric area, or gap g between these two sets represents the *metabolic error* of the individual (see figure I.2). It takes into account both the over- or under-realization of each biological process. The smaller this metabolic error, the better the individual (*i.e.* the higher its fitness) and the greater its probability of reproduction. Ultimately, an individual with a phenotype matching precisely the environmental target would be considered perfect. However, as the environment is defined as a non-linear continuous function, improving the phenotypic approximation of the phenotype can only be done by increasing the number of proteins to infinity.

5 From Genotype to Phenotype

In Aevol, the way the genotype is decoded into a phenotype is directly inspired from how it is achieved in bacteria. It basically follows the central dogma of molecular biology (Crick, 1970), the genetic information flowing from DNA to proteins through RNA in a transcription-translation process, and the primary sequence of proteins determining their metabolic activity through what we could consider a "folding" process. We basically defined a set of signalling sequences that allow us to identify the sequences that will be transcribed into RNAs and, within these RNAs, the sequences that will be translated into proteins. These proteins will then be interpreted in terms of realized or inhibited biological processes.

5.1 Transcription: from DNA to RNAs

In bacteria, transcription initiates at particular sites, the promoters, where an RNA-polymerase can recognize a consensus sequence and bind to the DNA to begin the synthesis of an RNA molecule. In Aevol, a promoter is a sequence whose Hamming distance d with a predefined consensus sequence is lower than or equal to d_{max}. The sequence we typically [1] use in our experiments is 22 base-pairs long: 0101011001110010010110 and we allow up to $d_{max} = 4$ mismatches. This sequence is long enough for non-coding sequences to have only a low probability of becoming a promoter after a mutation.
When a promoter is identified, the following sequence is transcribed until a terminator is found. Terminators must be more frequent than promoters to limit the overlapping of transcribed sequences. Using a consensus based signal for terminators would therefore require that consensus to be short. Then, this particular sequence could no longer be present in any gene within the model, applying a strong constraint on gene sequences. Interestingly, most bacterial terminators are not based upon a consensus, rather, they follow a general pattern that renders a hair-pin structure because of the base-complimentarity property. Remarkably, these ρ-independent terminators have the property of being both long *and* frequent. We therefore defined a terminator as a sequence able to form a hairpin, usually of length 4 for the stem and 3 for the loop. A terminator is then a sequence following the pattern $abcd***\bar{d}\bar{c}\bar{b}\bar{a}$ where $a, b, c, d = 0$ or 1 and $*$ is a wildcard.
A transcribed sequence (RNA) in Aevol is thus identified by a promoter-terminator couple. However, if one promoter marks the beginning of only one RNA, a terminator can mark the end of several. Indeed, when several promoters follow one another with no terminators in between, some DNA sequences can be transcribed onto several different RNAs. It is also worth noting that terminators form a hair-pin on both the leading and the lagging strands, marking the end of transcription on both strands while promoters can only initiate transcription on one strand.
The expression level e of each RNA depends on the "quality" of its promoter; the closer its sequence to the consensus, the higher the expression level, namely $e = 1 - \frac{d}{d_{max}+1}$. This modulation of gene expression models in a simple way the interaction between the RNA-polymerase and the promoter without explicit regulation. As we will see in chapter III,

1. All the consensus signals are defined as parameters of the model and hence can be modified. Throughout this work, we used the same signals as in (Knibbe et al., 2007a).

an extension of the model (R-Aevol) precisely allows for experiments with gene regulation networks.

5.2 Translation: from RNA to Proteins

Similarly to the transcription process, the translation process is driven by signalling sequences. These signals are searched for on the transcribed sequences but while transcription initiation signals were searched for on both strands (in opposite directions), translation initiation signals are only searched for in the promoter-to-terminator sense. The gene initiation signal consists of a Shine-Dalgarno like sequence (usually 011011) followed, a few base pairs further, by a Start codon (000 - see table I.1). In previous experiments, the spacer between a Shine-Dalgarno sequence and a START codon was set to 3, however, this particular number representing exactly the size of a codon, it was set to 4 in all further experiments. Hence, any sequence following the pattern $011011****000$ within an RNA marks the beginning of a gene. Whenever an initiation signal is detected, the following sequence is translated three bases (one codon) at a time until a Stop codon (001) is found on the same reading frame. Each of these codons is translated into the corresponding amino-acid in the genetic code (table I.1), thus forming the primary sequence of the protein that is being synthesized. Note that the genetic code is not redundant, meaning that there is no bias towards any particular Amino-Acid.

5.3 Computation of the Metabolic Activity of a Protein

As seen in section 4, a protein's metabolic activity is represented in our fuzzy set environment framework as a triangle characterized by its mean value m, height h and half-width w (figure I.1-f). The primary sequence of a protein is broken down into a set of three interlaced variable-length binary strings, corresponding to each of these parameters (figure I.1-e), following the genetic code (table I.1). For example, the codon 100 (resp. 101) which is translated as a $M0$ (resp. $M1$) amino-acid, contributes to the m parameter by adding a 0 (resp. 1) bit to its binary code. The binary sequence corresponding to each parameter is then interpreted as an integer using the Gray binary code (a code in which 2 successive values differ only by one bit) and finally normalized into the interval of definition of the parameter according to the length of the sequence ($m \in [0,1], h \in [-1,1]$ and $w \in [0, w_{max}]$, w_{max} being a parameter of the model). The longer the sequence, the more precise the value of the parameter. The resulting fuzzy set represents the efficacy with which each metabolic process in $[m-w; m+w]$ is realized (if $h > 0$) or inhibited (if $h < 0$), with a preference for those processes closest to m for which the maximum possibility degree of $e.|h|$ is achieved.

5.4 Phenotype Computation

Once all the proteins have been characterized, the phenotype can be computed by combining each of their metabolic contributions. The phenotypic framework of Aevol allows for both pleiotropy, a single protein being able to contribute to different biological processes, and polygeny. Indeed, as several proteins can contribute to the same biological process, the degree of possibility associated to this process (*i.e.* the efficacy with which

the organism will be able to realize it) will then result from the interaction of these proteins. To combine the activity of all the proteins, now represented as fuzzy sets, we used Lucasiewicz's operators (equation I.1), that are better suited for modelling interactions than most fuzzy operators.

$$\begin{cases} NOT : \overline{x} = 1 - x \\ AND : x \cap y = min(x + y, 1) \\ OR : x \cup y = max(x + y - 1, 0) \end{cases} \quad \text{(I.1)}$$

The organism's phenotype is then given by $P = (\cup A_i) \cap (\cup \overline{I_j})$, with A_i the phenotypic contribution of the i-th activating protein and I_j that of the j-th inhibiting protein. This phenotype represents the efficacy with which the organism realizes each biological process in Ω.

6 Selection and Evolutionary Loop

When the phenotype P of an individual has been computed, it is compared to the environmental target E. The area, or *gap* g between these two fuzzy sets is the "*metabolic error*" of the organism, that quantifies how far it is from being perfectly adapted to its environment. Once this "metabolic error" is known for all the individuals in the population, a probability of reproduction is assigned to each organism according to either its rank in the population or directly to this metabolic error value. Three different selection schemes have been implemented in Aevol.

In the *fitness proportionate* scheme, the probability of reproduction of each organism is proportional to its fitness. A simple proportion to $1/g$ would produce a very mild selection, yielding an evolution based almost exclusively on random drift, the probability of reproduction is hence proportional to $exp(-k.g)$, where k influences the selection intensity.

The other two selection schemes are based on the rank of the organisms in the population, which allows us to maintain a constant selective pressure throughout the entire evolutionary process. Organisms are thus first sorted by increasing fitness (the worst individual in the population having rank 1). Then, their probability of reproduction can be computed depending on their rank r and according to whether the linear or exponential scheme is used.

For the *linear ranking* scheme, the probability of reproduction of an individual is given by $p_{reprod} = \frac{1}{N}.(\eta^- + (\eta^+ - \eta^-).\frac{r-1}{N-1})$, where $\frac{\eta^+}{N}$ and $\frac{\eta^-}{N}$ represent the probability of reproduction of the best and worst individual respectively. The population size being fixed, η^- must be equal to $2 - \eta^+$. As for η^+, it must be chosen in the interval $[1, 2]$ so that the probability increases with the rank and remains in $[0, 1]$.

For the *exponential ranking* scheme, the probability of reproduction is given by $p_{reprod} = \frac{c-1}{c^N-1}.c^{N-r}$, where $c \in]0, 1[$ determines the intensity of selection (the closer to 1, the weaker the selection).

Whichever selection scheme is used, for each of the N (size of the population) offspring that will be produced, its actual parent is randomly drawn through a roulette wheel process, the roulette wheel being biased according to the probability of reproduction of each organism. The offspring will receive a copy of the parental chromosome after this chromosome has undergone a process of chromosomal rearrangements and local mutations. The newly

formed population replaces the former one and the next iteration of the evolutionary loop can begin. In the last chapter of this manuscript, an extension of the model is presented in which a horizontal transfer event can be involved during the replication process.

7 Genetic operators

During their replication, genomes can undergo different kinds of modifications acting on different scales: some (switches and indels) act locally, modifying the sequence intrinsiquely while others (chromosomal rearrangements) act on a larger scale, changing the organization of the sequence.

7.1 Local Mutations

Single Base Substitution (Switch)

This is the most basic mutation that is modelled in Aevol. A single base in the sequence is switched from 0 to 1 or from 1 to 0.
When this type of mutation happens in a non-coding region, it is most likely silent. However, when it happens in a coding sequence, it has the potential of destroying a signalling sequence, for instance a transcription initiation signal (promoter), hence destroying the corresponding RNA and the putative genes it carried, or a translation termination signal (Stop codon), hence lengthening the corresponding gene (or rendering it non-functional if there is no other Stop codon before the terminator).
Of course, a switch can also create a new signalling sequence, however, this seldom leads to the creation of a new gene. Indeed the prerequisites for creating a new gene from a random sequence are stringent: there must be one signalling sequence of each kind and in the right order (promoter, Shine-Dalgarno-Start, Stop codon, terminator). Furthermore, there is no selection to direct the genetic drift towards such a sequence: it either appears as a whole or has no phenotypic contribution whatsoever.
Finally, when a switch occurs within a gene, it modifies one single codon of the gene, *i.e.* one single amino-acid of the corresponding protein's primary sequence. Such a modification can modify one characteristic of the phenotypic triangle (or two if the new codon is not of the same kind as that of the former one, *e.g.* an $M0$ becoming an $H0$) either slightly or substantially depending on local conditions. Mutations having a mild impact on the phenotype allow for gradual modifications of the protein through successive mutations within a lineage.

Indels

Small insertions or small deletions consist of inserting a few exogenous nucleotides into the genome or deleting a few nucleotides from the genome (typically up to 6). These mutations can have the same effects as single base mutations with regard to the creation or destruction of signalling sequences. However, when occurring inside a gene, a distinction must be made. Indeed, while an indel of 3 or 6 bases within a gene is likely to have only a limited effect on phenotype, inserting or deleting a number of bases that is not a multiple of 3 can cause a frameshift mutation, thus modifying the whole sequence of the gene

downstream from the mutation and most likely changing its length. Indeed, as nucleotides are translated three at a time until an in-frame STOP codon is found, inserting or deleting *e.g.* two nucleotides will cause the whole sequence of the gene to be read in a different frame, including the STOP codon that will thus not be identified as such.

Indels are hence mutations which, although they act locally, can still have drastic effects on a scale of up to a few genes.

7.2 Chromosomal Rearrangements

Chromosomal rearrangements are large scale events that can involve sequences of any size on the chromosome (up to the whole genome altogether). In bacteria, several mechanisms can result in a rearranged chromosome, including double strand break (DSB) repair mechanisms or DNA polymerases "jumping" from a sequence to another during replication. Depending on the localization and direction of the involved sequences, such events can lead to different kinds of rearrangements: duplications, deletions, translocations and inversions (Higgins, 2005; Lewin, 2007).

- Duplication: a randomly chosen segment is duplicated, the copy being reinserted either side by side with the template (tandem duplication) or at a random position on the genome [1].
- Deletions: a randomly chosen segment is deleted from the genome.
- Translocations: a randomly chosen segment is excised from the genome. The segment is then circularized and reinserted at a random position.
- Inversions: a randomly chosen segment is inverted, the sequence on each strand switching both strand and direction.

Each kind of rearrangement has different consequences: inversions and translocations are quite conservative since they basically consist in moving sequences around. The only changes that can have a direct influence on the phenotype are at the breakpoints, where the sequences are cut, potentially breaking genes or RNAs, and then put back together, potentially creating new ones. Thus, even though they can lead to massive reorganizations of the genome, the impact of inversions and translocations on the content of the genome is similar to that of local mutations. As for duplications and deletions, their effect can be dramatic since they can cause a whole segment of the genome to be irrevocably lost or a group of genes to be fully duplicated. Ultimately, duplications and deletions can greatly modify the size and the proportion of coding sequences of the genome.

Most of the mechanisms that can cause a chromosomal rearrangement depend on the presence of similarities in the genetic sequence. In the standard version of the model however, alignments are not mandatory for a rearrangement to occur, breakpoints being simply drawn in a uniform distribution in $[0; L[$ where L is the genome length. One of the main objectives of this work was to develop a more precise model of chromosomal rearrangements in which rearrangements are more likely to occur at breakpoints that are similar in sequence. This extension of the model is presented in Chapter IV.

1. In the former version of the model, tandem duplications are not formally modelled, the duplicated segment being reinserted at a random point along the chromosome.

8 *In Silico* Experimental Evolution with Aevol

Digital genetics models in general, and Aevol in particular, are very complex systems, meaning that it is impossible to conduct a formal analysis on any such model. Besides, simulations usually take a lot of time (up to a few months of calculation in extreme cases), making it inconceivable to explore a vast parameter space. One could then argue that there is no interest in building such a complex model. However, while such a model is indeed complex, it is nonetheless a lot simpler than the initial object or process, making it easier to study. Also, while simulations can take quite some time, the time scales are fundamentally different and even a long digital genetics experiment is several orders of magnitude faster than an evolutionary experiment using real organisms. Furthermore, models allow the practitioner to have better control of the conditions in which an experiment is conducted and also a complete and exact fossil record of the whole evolutionary process. As Richard Lenski said, "There are no missing links in the digital world" (cited in O'Neill (2003)). Finally, digital genetics make it possible to conduct experiments that are impossible to realize with real organisms (O'Neill, 2003).
In any case, the use of digital genetics models remains an experimental approach that thus requires a strict experimental method. Therefore, the typical use of a digital genetics model is very close to "wet" experimental evolution procedures (Elena and Lenski, 2003). Populations of organisms are initialized and left to evolve in controlled conditions. Usually one, and at most a few parameters are set to different values from one simulation to the next. Then, by observing the products and dynamics of the evolutionary process in the different conditions tested and by comparing them, one can shed light onto the impact of the tested parameters and eventually unravel the direct and indirect pressures that constrain the structure of the organisms. Table I.2 presents the parameters of the Aevol model. Note that several simulations are needed for each set of parameters to assess the repeatability and the statistical significance of the observations, but this can be easily achieved on computer farms or clusters.

At the end of a simulation, the line of descent (lineage) of the best individual of the last generation can be reconstructed from the log files. One can then analyse *e.g.* the particular mutational events that went to fixation.
Despite its numerous advantages, *in silico* experimental evolution is still very demanding in terms of computational resources. That is the reason why the model was thoroughly reimplemented during this PhD. A great deal of effort has been dedicated to making the code modular and maintainable while optimizing resource usage, in particular memory and execution time (the latter being reduced 10-fold compared to earlier versions). It is now possible to conduct large scale experiments, testing a wide combination of parameters for hundreds of thousands of generations. This new implementation has also made it possible to run large scale experiments using extensions of the model such as R-Aevol or the alignment-based extension, that are a lot more computationally costly than the core model. In fact, this new implementation is what made this work possible.
Some limitations remain, however, in particular regarding the size of the population we can simulate. The memory of one single computer is indeed limited and resorting to disk space to virtually increase it would come with an unacceptable increase of computational

Parameter	Meaning
N	Population Size
nb_gener	Number of Generations to be run
$init\_length$	Size of the initial, randomly generated genomes
$init\_method$	Initial Population Generation Method. Whether to use a bootstrap (generate genomes until the corresponding phenotype is better than the "flat" phenotype) and whether to initialize the population with clones of a single organism or with different organisms
$selection\_scheme$	Selection scheme to use (linear ranking, exponential ranking or fitness proportionate)
k, η^+ or c	Selective pressure (see section 6)
$E = \sum_i \alpha_i G_i$	The environment is defined as the sum of any number of Gaussian curves
$env\_sampling$	The environment is discretized as a piecewise-linear fuzzy set this is the number of points to generate from the Gaussians
μ_{point}	Point Mutation Rate
μ_{s_ins}	Small Insertion Rate
μ_{s_del}	Small Deletion Rate
μ_{dupl}	Large Duplication Rate
μ_{del}	Large Deletion Rate
μ_{inv}	Large Inversion Rate
μ_{trans}	Large Translocation Rate
max_indel_size	Maximum number of bases inserted or deleted by an indel
W_{max}	Maximum Pleiotropy of Proteins

Table I.2 – Main parameters of the Aevol model. Parameters that are specific to a particular extension of the model will be presented along with the extension.

time. We are hence looking forward to implementing a parallel version of the model, that will allow for both a faster execution (useful for demanding extensions of the model) and larger population sizes (thanks to distributed memory).

Even though Aevol is under constant evolution, it is now stabilized and its use no longer requires a great expertise of the model itself. Experiments can be conducted simply by changing the parameters of the model. Aevol is available online as both stable releases and a *subversion* (SVN) repository [1]. This allowed in particular for an active collaboration with a team from the INSERM in Paris that have extended the model to study the emergence of cooperation (current versions of the model allow to use various extensions of the model with the same code, depending on the parameters). We are now planning to initiate a collaboration with the York Center for Complex Systems Analysis.

1. http://gforge.liris.cnrs.fr/projects/aevol/

9 Seminal Results

Aevol was designed by Carole Knibbe and Guillaume Beslon. During her PhD thesis, C. Knibbe explored the global properties of the model, discovering a strong second-order pressure that – at least in the model – strongly determines several structural characteristics of the genome such as the genome size, the number of genes and the amount of non-coding DNA.

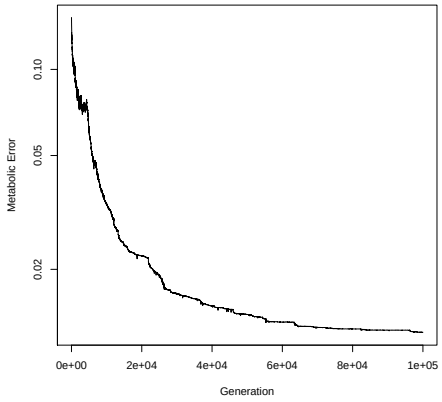
9.1 A Typical Run in Aevol

Using Aevol, full campaigns of evolution experiments can be run for hundreds of thousands of generations in a very reasonable time (a few days for 500,000 generations of a population of 1,000 individuals with default parameters). One can then conduct an analysis of the different genomic structures obtained with different parameters and propose hypotheses to explain the observed patterns. However, although the final genomic structures can be very different from one set of parameter to the other, the evolutionary process is itself relatively stable over most simulations.

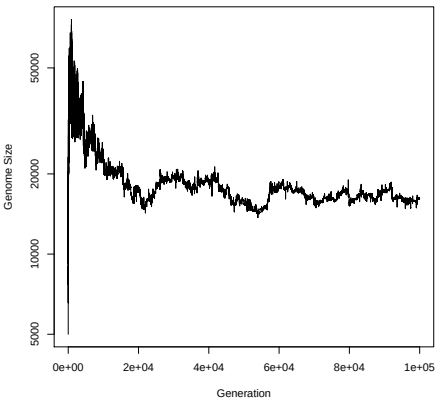
As the initial population is usually filled with clones of a randomly generated individual [1], the potential of adaptation is very high at the beginning of a simulation. Hence, during the first few hundred generations, the organisms can improve very quickly, adapting to their new environment. This first step of evolution is mainly driven by the recruitment of new genes through gene duplication-divergence. Because at this stage of evolution, duplications are often beneficial, allowing the organisms to literally fill the gap between the initial phenotype and the environmental target by creating new genes, this first burst of adaptation comes along with an explosion of the size of the genome (figure I.3), that increases from 5,000 base-pairs (bp – default initial size) to up to hundreds of thousands of bp. In fact, every structural parameter (genome size, number of genes, overall size of non-coding sequences) follows this pattern of extremely fast increase. This first stage of evolution is immediately followed by a very fast decrease in the size of the genome, accompanied by a loss of genes while the fitness continues to get better. Finally, during the third stage of evolution, the genome stabilises in size and the number of genes it bears increases slowly while they continuously tend to get longer, gaining in precision.

In Aevol, organisms can improve their fitness by acquiring new genes and/or by improving those they already own. The progress of evolution shows that these mechanisms are not used equally during the whole evolutionary process: organisms first enlarge their genome and gene repertoire, usually recruiting new genes by duplication-divergence of existing ones. Then, the coding sequences of these genes are improved while gene recruitment continues, though, at a lower rate. Note that fitness is never completely stable and that beneficial mutations occur regularly throughout the whole evolution. Even at advanced stages of evolution, selection thus remains directional.

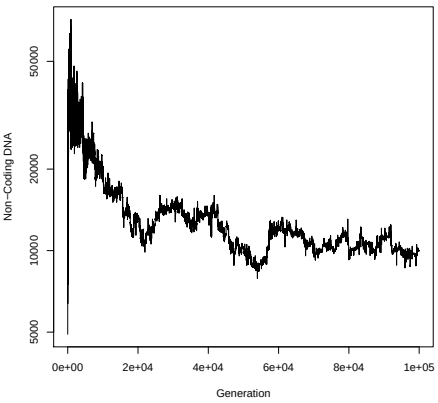
1. A new random individual is generated and evaluated until its phenotype is better than the "flat" phenotype.



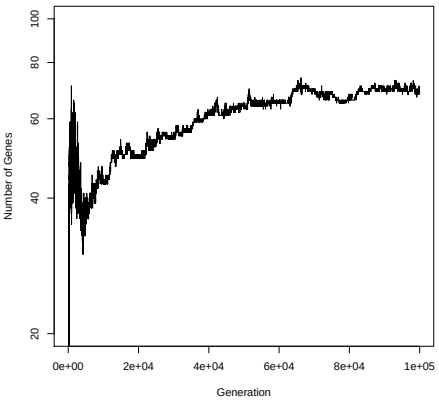
(a) Metabolic error (fitness measure).



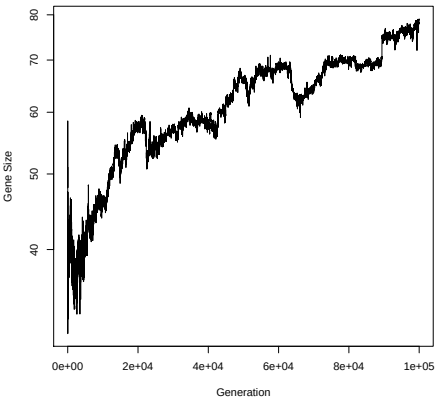
(b) Genome size.



(c) Amount of non-coding DNA.



(d) Number of genes.



(e) Average gene size.

Figure I.3 – Evolution of the fitness and of a few genomic characteristics of the best individual of each generation in a typical run.

9.2 Evolution of Evolvability

As we mentioned in the introduction of this manuscript, experiments using the Aevol model have shed light on a very interesting second-order pressure. Specifically, it was shown that the rate of mutations in general, and rearrangements in particular, is a strong determinant of the size and structure of the genome. It was also shown that this is due to a long term selection of a specific level of mutational variability of the phenotype.
During her PhD thesis, C. Knibbe allowed 72 populations of 1,000 individuals to evolve during 20,000 generations under different mutation rates and selective pressures. The resulting genome sizes and the number of genes they bore clearly scaled as a power law of the mutation rate for each intensity of selection. This scaling of the genome size with respect to the mutation rate is very similar to what was observed in DNA-based microbes (Drake, 1991). Interestingly the amount of non-coding sequences in the genomes also scaled as a power-law of the mutation rate, furthermore with a greater exponent than coding sequences (figure I.4). Knibbe et al. (2007a) have shown that this is due to the mutagenic effect of the non coding sequences on the surrounding genes when it comes to large duplications and deletions: while local modifications of the sequence such as point mutations or indels can be assumed to have no effect on the phenotype when they occur in non-coding sequences, duplications and deletions, even when their breakpoints are in non-coding sequences, can duplicate or delete huge portions of the genome, including genes. Non-coding sequences can hence be considered as a passive substrate promoting chromosomal rearrangements, thus widening the mutational target to the entire chromosome when it could be thought to be limited to coding sequences.
A striking observation was that, for each intensity of selection, the fraction of neutral offspring (*i.e.* the fraction of organisms having the same phenotype as their parent) F_ν of the best individual in the last generation was roughly the same whatever the mutation rate (figure I.5). In fact, the evolved $F_\nu W$ (with W, the number of reproductive trials of the best individual in the population) was always close to the value that, given the selective pressure, would tend to make the best individual produce one single neutral offspring, the rest of its progeny undergoing changes in their phenotype. This suggests the long-term selection of a particular trade-off between exploration and exploitation, or in other words between the maintenance of the ancestral phenotype and the search for better ones.

9.3 Conclusion

Aevol is hence a digital genetics model in which the structure of the genome is very realistic and is free to evolve. It integrates central genetic features and mechanisms and in particular intermediate levels of organisation between the genome and the phenotype (the transcriptome and the proteome) as well as realistic operators for both mutations and rearrangements. Experiments have shown that, in Aevol, organisms are selected on the basis of both their direct adaptation value and of indirect criteria such as the level of mutational variability of their phenotype. Aevol is hence particularly suited for the study of genome organization as a result of second-order selective pressures.

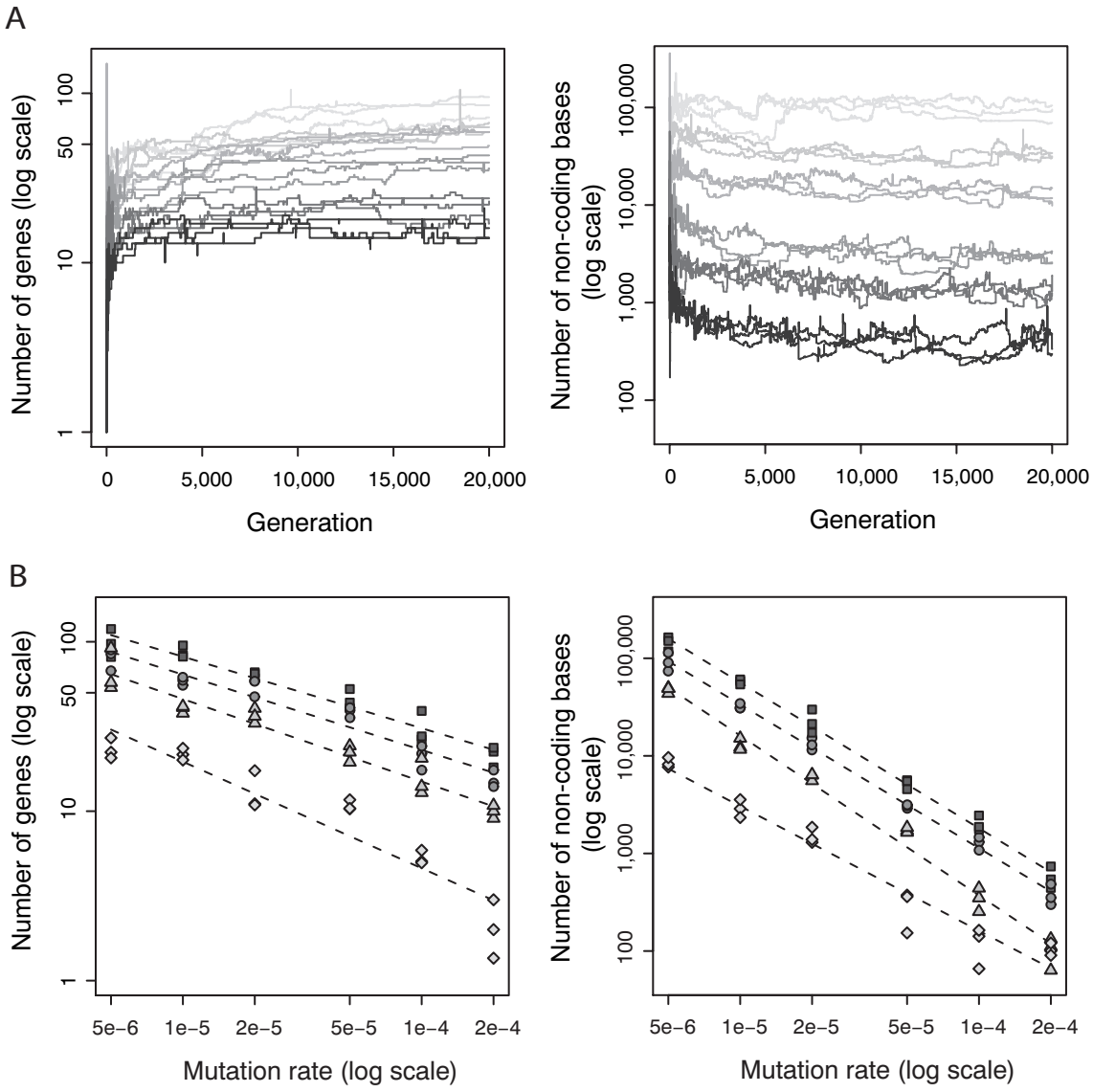


Figure I.4 – (from Knibbe et al., 2007a). **(A)**: Evolution of the number of genes and of the amount of non-coding sequences in the line of descent of the final best organism of each simulation with an exponential ranking selection scheme and $c = 0.995$. The common rate μ_{mr} for mutations and rearrangements ranges from 5×10^{-6} (light grey) to 2×10^{-4} (black) per type of mutation per base pair per replication. Both the number of genes and the amount of non-coding sequences clearly depend on the mutation/rearrangement rate, the lower μ_{mr}, the more genes and non-coding bases. **(B)**: Number of genes and amount of non-coding sequences of the final best individual of each simulation. For the four values of selective pressure tested ($c = 0.9900$: squares, 0.9950: circles, 0.9980: triangles, and 0.9995: diamonds), both the number of genes and the amount of non-coding sequences scale as a power law of μ_{mr}.

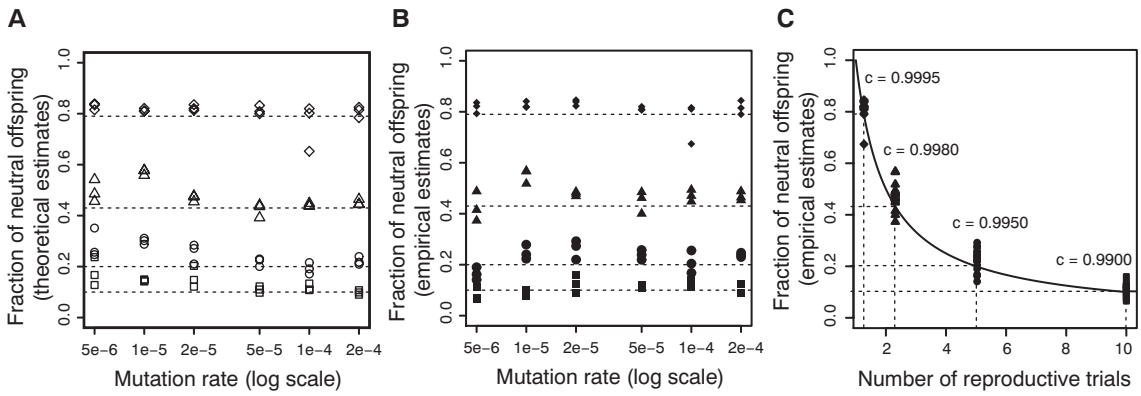


Figure I.5 – (from Knibbe et al., 2007a). The intensity of the selection sets the appropriate level of variability. For each run, the fraction F_ν of neutral offspring of the final best organism was estimated, both theoretically (**A** – see (Knibbe et al., 2007a)) and empirically (by simulating 1,000 independent replications: **B**). **(A and B)**: for a given selection intensity ($c = 0.9900$: squares, 0.9950: circles, 0.9980: triangles, and 0.9995: diamonds), this evolved F_ν is roughly the same for the six mutation rates tested despite the huge diversity of genome structure. The evolved F_ν is close to the value that would ensure an average of one neutral offspring to the best individual (dotted horizontal lines). **(C)**: the evolved F_ν as a function of W, the number of reproductive trials of the best individual in the population. They are indeed close to $1/W$ (black curve).

Chapter II

Indirect Selection and the Organization of Transcription

The results presented in this chapter have been partly published in Parsons et al. (2010b), based upon a smaller set of data where the organisms evolved during 20,000 generations in a less demanding environment.

1 Introduction

In the previous chapter, we presented results obtained with the Aevol model, which allowed us to identify a very strong second-order selective pressure towards a specific level of mutational variability of the phenotype. Consequences of this pressure on the genome structure include a strong trend for organisms having evolved under low mutation rates to have a very large genome with many genes and a huge proportion of non-coding sequences. On the contrary, organisms having evolved under very high mutation rates have very short genomes containing fewer genes and a very small proportion of non-coding sequences. These effects were shown to be mostly due to the rates at which chromosomal rearrangements occur, particularly duplications and deletions, that affect vast areas of the genome. Chromosomal rearrangements can occur with breakpoints in any part of the genome, regardless of whether these breakpoints fall in either coding or non-coding regions. Since large duplications and deletions impact not only the sequences around the breakpoints but the whole region between the breakpoints, this kind of event can affect genes (either duplicating or deleting them) even when all the breakpoints are in non-coding sequences. Thus, when large duplications and deletions are involved, non-coding regions have a mutagenic effect on the surrounding genes. Then a population of organisms owning a large genome, even mostly non-coding, in a context of high rearrangement rates, would undergo so many rearrangements that it would fall into error catastrophe, *i.e.* it would be unable to maintain its phenotype. On the contrary, organisms owning a very

small genome in a context of low rearrangement rates would produce very few offspring differing from their parents, yielding a very poor evolvability.
Consequences of this second-order selective pressure on other levels of organization than the genome are of the utmost interest, especially when it comes to the structure of the transcriptome. Indeed, when one looks at real organisms, differences in the size and structure of the genome usually come along with great variations in the way the genome is transcribed. On the one hand, short genomes are usually almost entirely transcribed, the resulting RNAs being most of the time quite long and containing several genes (polycistronic RNAs). In extreme cases, the whole genome can be transcribed in only a couple of RNAs (Zheng and Baker, 2006). On the other hand, long genomes usually give rise to short RNAs (after splicing), very few of which contain more than one single gene and most containing no genes at all. These non-coding RNAs have received a great deal of attention in the last few years (Ponjavic et al., 2007; Will et al., 2007), in particular micro-RNAs that are thought to play a major role in the regulation of gene expression (Mattick and Makunin, 2006; Kapranov et al., 2007).
What mechanisms are responsible for these variations in the organization of transcripts and their relative importance remain open questions. Most efforts in these matters have been focused on understanding the evolution of operon structures. Operons are very interesting DNA structures where several coding sequences (often functionally-related) are packed together and transcribed together on a single messenger RNA (mRNAs). Operons have been the subject of a great number of studies resulting in a set of theories that try to explain their assembly and maintenance. The following summarizes the most defended of these theories:

The coregulation model is the original theory that came along with the discovery of the operon structure (Jacob et al., 1960). It claims that packing several functionally related genes together on the same mRNA is beneficial because they share their regulation sites, which means that variations of the transcription level (either because of mutations on the promoter or because of regulation) will preserve the relative expression levels of the gene products. According to this hypothesis, genes within an operon should be likely to be functionally related.

The selfish operon theory postulates that clustering genes for weakly selected functions together is beneficial for the genes themselves as it allows them to be horizontally transferred as a whole (fully functional unit), hence conferring a better advantage to the receiver than they would have provided individually (Lawrence, 1999). In the light of this theory, horizontal transfer is a necessary condition for the emergence of operons, which should contain preferentially genes that are functionally related.

The mutational burden theory, finally, propounds that it is the mutational hazard that constrains the total amount of DNA: The larger the amount of excess DNA (intergenic DNA, 3' and 5' UTRs, ...), the higher the probability of a rearrangement to occur within it, potentially inactivating coding sequences or else disturbing the dynamics of existing genes. Following this idea, a population subject to high rearrangement rates will face pressure to make genomes denser (Lynch, 2006; Knibbe et al., 2007a). In some cases, this densification may reach a point where transcribed regions can actually merge or where a transcribed region can contain several trans-

lated sequences thus composing an operon. In extreme situations, genes can even share a part of their sequence and overlap. Both merging transcribed regions or making genes overlap further reduces the size of the mutational target of the phenotype. This second order selective pressure for "streamlining" makes no assumption regarding gene function or horizontal transfer, so operons should be able to arise in the absence of transfer, putting together genes "working together" as well as functionally unrelated genes. In this view, the presence of operons must depend on the rearrangement rates, the selection strength and the population size.

Each of these theories has received evidence both for and against it. For instance, Pál and Hurst (2004) argue that the gene composition of operons in *E. Coli* is incompatible with the selfish operon theory but Hershberg et al. (2005) and Rensing (2002) suggest that it can explain at least some operon structures. As a matter of fact, it is very difficult to validate any of these models since the underlying processes are complex and act on a very long time scale. We propose here to investigate the organization of transcripts using the Aevol model.

2 Experimental Setup

We used Aevol to allow 245 populations of 1,000 individuals to evolve independently for 50,000 generations in near identical conditions where the only changing parameters were the mutation rate and the rearrangement rate (one common rate μ_m for each kind of local mutations, *i.e.* small insertions, small deletions and point mutations, and one common rate μ_r for each type of chromosomal rearrangements, *i.e.* duplications, deletions, inversions and translocations). We tested all the combinations of 7 different values (10^{-6}, 2×10^{-6}, 5×10^{-6}, 10^{-5}, 2×10^{-5}, 5×10^{-5} and 10^{-4}) for μ_m and μ_r, each combination being repeated five times with independent pseudorandom number generator seeds, yielding among other features a different initial population and different mutational events. The selection scheme that was used in all these experiments is the exponential ranking scheme, with a selection pressure set to 0.998. Finally, the environment we used is strictly the same as the one used for the experiments presented in chapter I, section 9.2 (see figure I.2). The complete set of parameters used in these experiments is presented in table II.1.

This experiment was designed as a null-experiment for the selfish operon theory: the populations evolved in a strictly clonal framework where no horizontal transfer was allowed. According to the selfish operon theory, selfish operons should not be observed in such conditions. Thus, operons that would arise in our experiments could not be explained by the selfish operon theory and could hence find their roots in either the co-regulation theory or the mutational burden hypothesis. Mutation and rearrangement rates can be varied to test the mutational hypothesis while the co-regulation theory can be addressed by analysing the functional relatedness of genes organized in operons.

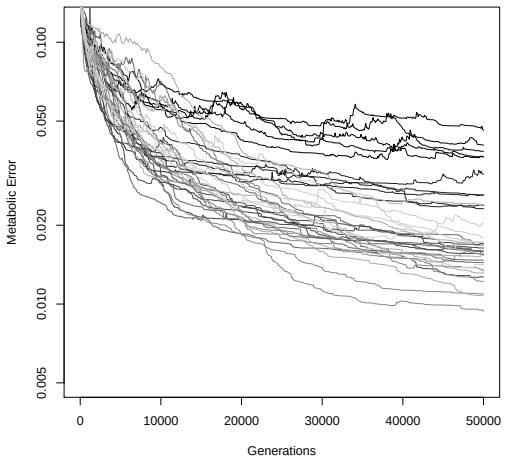
Parameter	Value
N	1,000
nb_gener	50,000
$init\_length$	5,000
$init\_method$	Clonal, One Good Gene
$selection\_scheme$	Exponential Ranking
c	0.998
$E = \sum_i \alpha_i G_i$	$\alpha_1 = 1.2; G_1 : \mu = 0.52; \sigma^2 = 0.12$
	$\alpha_2 = -1.4; G_2 : \mu = 0.2; \sigma^2 = 0.07$
	$\alpha_3 = 0.3; G_3 : \mu = 0.8; \sigma^2 = 0.03$
$env\_sampling$	300
μ_{point}	$\mu_m \in \{10^{-6}, 2\times10^{-6}, 5\times10^{-6}, 10^{-5}, 2\times10^{-5}, 5\times10^{-5}, 10^{-4}\}$
μ_{s_ins}	
μ_{s_del}	
μ_{dupl}	$\mu_r \in \{10^{-6}, 2\times10^{-6}, 5\times10^{-6}, 10^{-5}, 2\times10^{-5}, 5\times10^{-5}, 10^{-4}\}$
μ_{del}	
μ_{inv}	
μ_{trans}	
max_indel_size	6
W_{max}	0.01

Table II.1 – Parameters used in all the experiments of this chapter. Mutation and rearrangement rates take their values among those proposed, one common value for the three types of local mutations, and one common value for the 4 types of rearrangements.

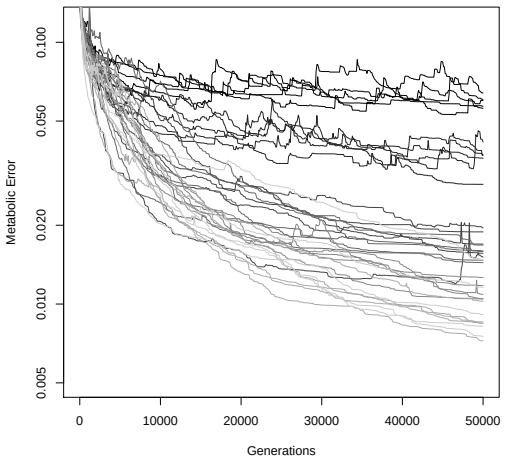
3 Results

3.1 Evolution of the Fitness

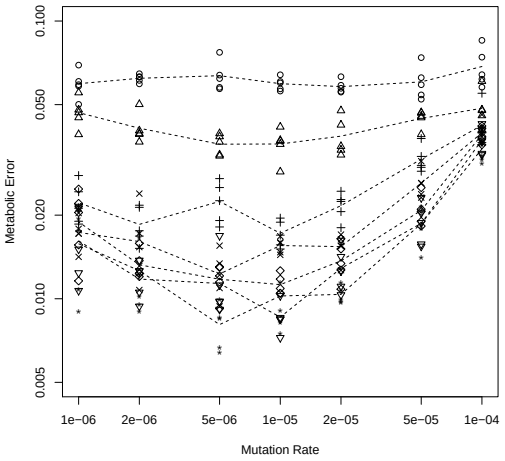
In these experiments, as opposed to those presented in (Knibbe et al., 2007a), two distinct rates were used for the genetic operators: one common rate μ_m for each kind of local mutation and one common rate μ_r for each kind of chromosomal rearrangement. This distinction allowed us to assess the predominance of the rearrangement rate in the effects of the indirect selective pressure we discussed in chapter I, section 9.2. The first striking result we observed is that the fitness strongly depends on both the mutation and the rearrangement rates (figure II.1). After 50,000 generations, the best organisms are those having evolved in a context of both low mutation rate and low rearrangement rate. Yet the rearrangement rate seems to have a greater impact on fitness than the mutation rate. In particular, when the rearrangement rate is very high ($\mu_r \geq 2 \times 10^{-4}$), the mutation rate seems to have very little effect on the final fitness (at least in the range tested). Note that, as shown in figures II.1(a) and II.1(b), this effect is not due to the non-convergence of some simulations.



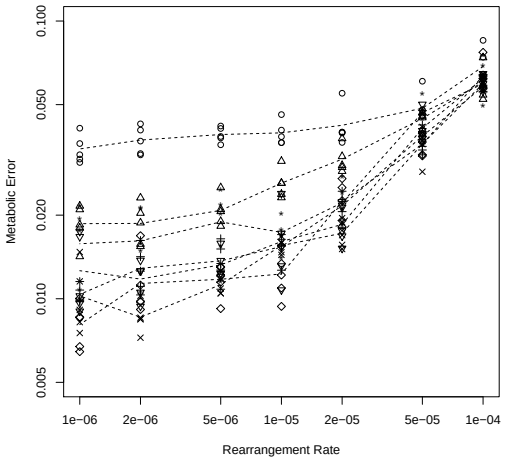
(a) Evolution of the metabolic Error



(b) Evolution of the metabolic Error



(c) Stabilized Metabolic Error



(d) Stabilized Metabolic Error

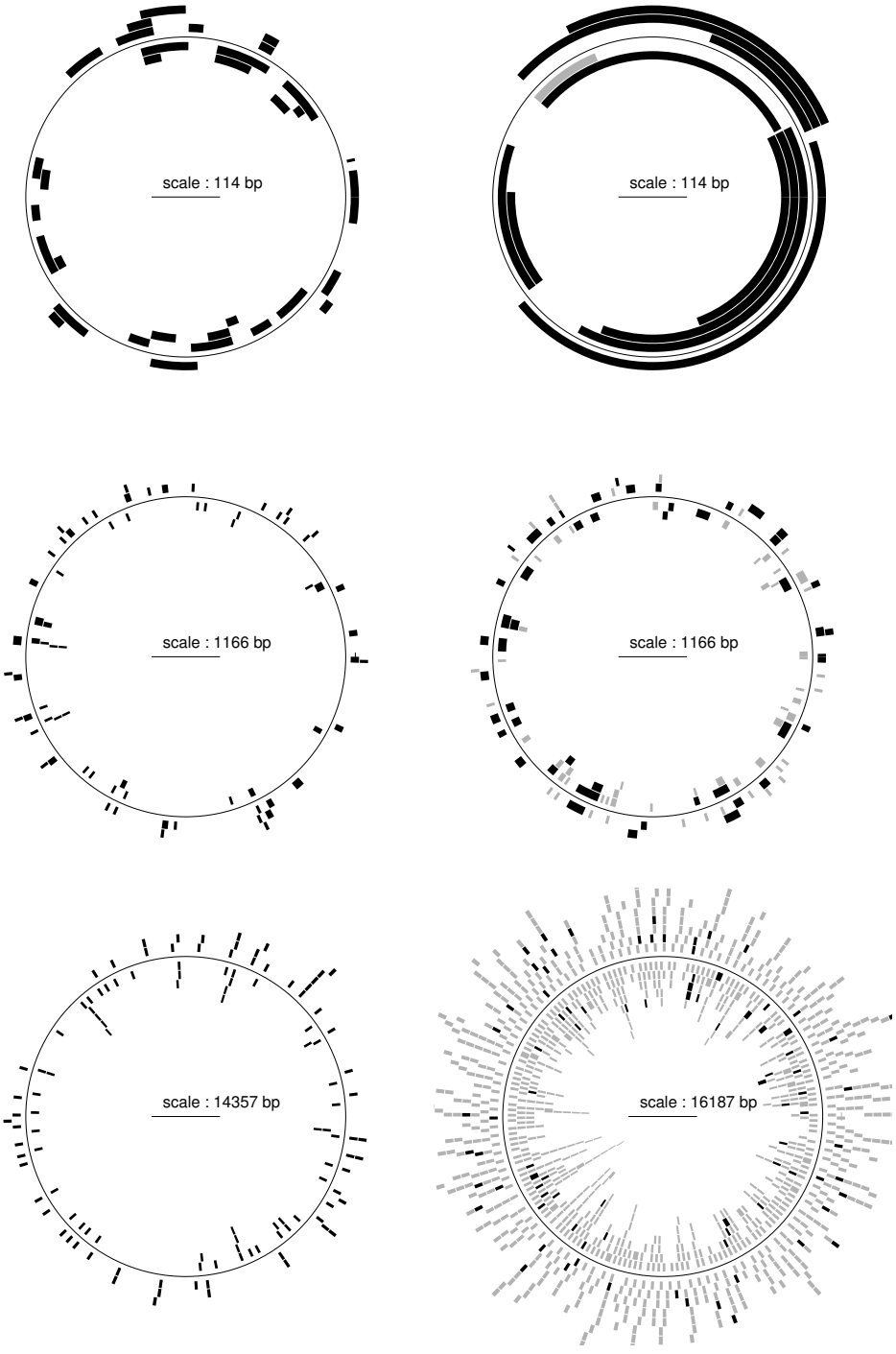
Figure II.1 – **(Top)**: evolution of the metabolic error in the lineage of the final best individual of each simulation with **(a)**: $\mu_r = 10^{-5}$ (different μ_m are represented by grey levels, black lines corresponding to $\mu_m = 10^{-4}$), **(b)**: $\mu_m = 10^{-5}$ (different μ_r are represented by grey levels, black lines corresponding to $\mu_r = 10^{-4}$).
(Bottom): metabolic error of the final best organism of each simulation, as a function of μ_m **(c)** and μ_r **(d)**. μ_r and μ_m are also respectively reported in **(c)** and **(d)** as the shape of the points (circles: 10^{-4}, upwards pointing triangles: 5×10^{-5}, plus signs: 2×10^{-5}, multiply signs: 10^{-5}, diamonds: 5×10^{-6}, downwards pointing triangles: 2×10^{-6}, stars: 10^{-6}). Both μ_m and μ_r have an impact on the metabolic error of the final best individual, the impact of μ_r being greater than that of μ_m. Organisms having evolved with the lowest μ_r (10^{-6}) and a low to medium μ_m ($< 2 \times 10^{-5}$) tend to be best adapted.

3.2 Evolution of the Structure of the Genome

After 50,000 generations of evolution with different rates of mutations and rearrangements, the evolved organisms have very different genome organization. According to previous

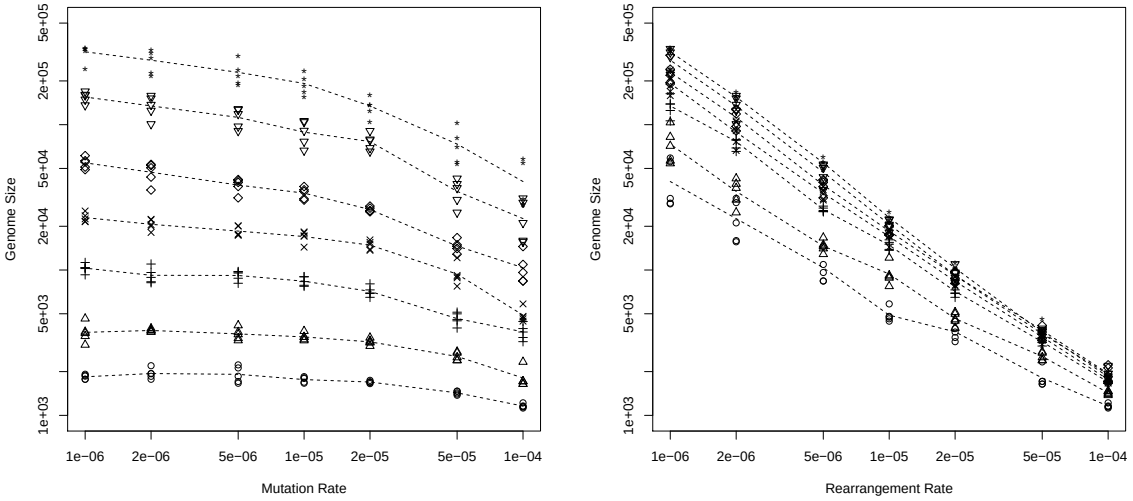
results obtained with the Aevol model (Knibbe, 2006; Knibbe et al., 2007a), the major determinant of genome size and structure should be the rearrangement rate. Figure II.2 shows the gene- and mRNA-content of the best organism of the last generation of three representative simulations with, from top to bottom, high (10^{-4}), moderate (10^{-5}) and low rearrangement (10^{-6}) rates, all three of them sharing the same moderate mutation rate (10^{-5}). These organisms are indeed very different: the higher the rearrangement rate they evolved with, the shorter and apparently the denser their genome.
Figure II.3 shows the genome size of the best individual of each simulation after 50,000 generations, respectively plotted versus the rearrangement rate and the mutation rate. According to these figures, the rearrangement rate is indeed the main determinant of the size of the genome although the local mutation rate also has an impact on it.
Interestingly, this effect is a lot stronger on non-coding sequences than it is on coding sequences. Figure II.4 shows both the amount of coding and non-coding sequences as a function of the rearrangement rate. The size of the non-coding sequences spans over four orders of magnitudes while the size of the coding sequences varies only 10-fold.
Indeed, considering the number of genes and the number of base pairs that are included in at least one gene on either strand (figure II.5), the predominance of the rearrangement rate over the local mutation rate seems to disappear. This points us directly to the *error threshold* effect (Biebricher and Eigen, 2005), according to which there is an upper bound to the mutation rate for a sequence of a given length (considered coding throughout its whole length) above which too many mutations are undergone, leading to *error catastrophe*: the sequence can no longer be maintained by the selection that is over-ruled by the mutational pressure. It is interesting to note that in the common understanding of the error threshold principle, this effect is confined to the coding sequences and is caused by point mutations. It is therefore not surprising to observe a clear effect of the local mutation rate on the coding sequences.
In our particular case, it is not the size of the sequence that is fixed but the per bp mutation and rearrangement rates; the size of the genome and the proportion of coding sequences are free to evolve. The size of the coding-sequences can hence adopt a size for which the genomic mutation and rearrangement rates are under the error threshold. This constraint on the size of the coding sequences appears to be quite mild under a value of around 2×10^{-5} per bp either for the mutation rate or the rearrangement rate. However, above this value, both the number of genes and the total amount of coding sequences collapse. This threshold is better seen on figure II.6 where we removed the points corresponding to the highest three values for μ_m when data is plotted vs μ_r and vice versa [1]. Considering figure II.1 in light of this observation, we can observe the same kind of variations, the loss of fitness being more drastic above this threshold of 2×10^{-5} for either μ_m or μ_r.
The mutation and rearrangement rates are hence the cause of a complex combined effect, the rearrangement rate impacting the whole genome (both coding and non-coding sequences) while the mutation rate only acts upon the coding sequences.
As shown by Knibbe et al., these effects are the result of an indirect selective pressure for a specific level of mutational variability of the phenotype: the size and structure of the evolved genomes are consistently such that the best individual in the population has on average at least one neutral offspring, *i.e.* one offspring having exactly the same

1. We removed these points because when the mutation rate and the rearrangement rate are both very high, their effects are strongly entangled so that it is difficult to differentiate their respective effects.

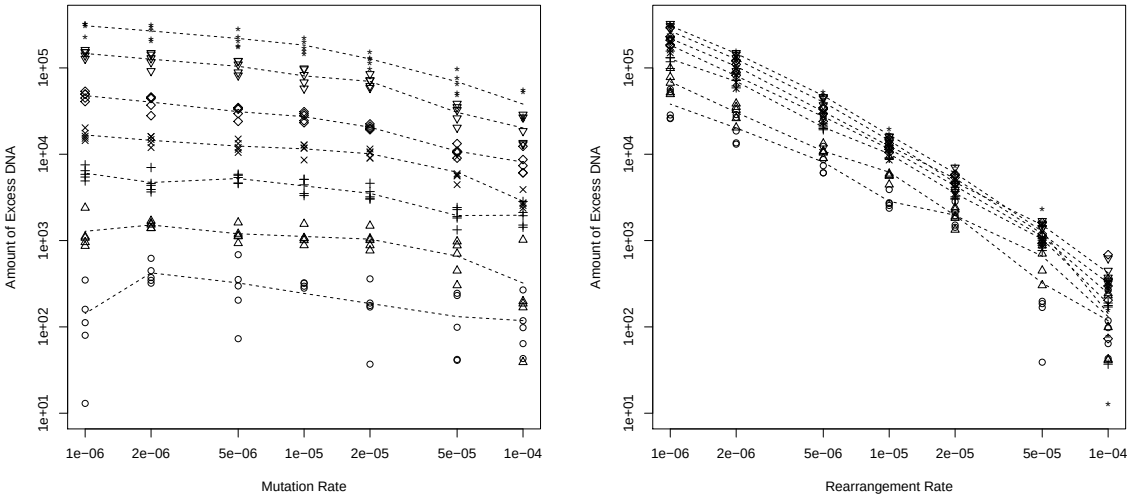


(a) Genomes with genes (b) Genomes with RNAs

Figure II.2 – Genome of the final best organism of typical simulations with respectively high (10^{-4} – top), medium (10^{-5} – middle) and low (10^{-6} – bottom) rearrangement rates. Left: black boxes represent genes. Right: black boxes represent coding RNAs (containing at least one gene), grey boxes represent non-coding RNAs.



(a) Genome size



(b) Amount of non-coding sequences

Figure II.3 – Genome size **(a)** and amount of non-coding sequences **(b)** of the final best organism of each simulation as a function of the mutation rate μ_m (left) and of the rearrangement rate μ_r (right). μ_r and μ_m are also reported as the shape of the points (circles: 10^{-4}, upwards pointing triangles: 5×10^{-5}, plus signs: 2×10^{-5}, multiply signs: 10^{-5}, diamonds: 5×10^{-6}, downwards pointing triangles: 2×10^{-6}, stars: 10^{-6}). Both μ_m and μ_r have an influence on the size of the genome and the amount of non-coding sequences but with a very strong predominance of μ_r.

phenotype as its parent. Figure II.7 shows the proportion F_ν of neutral offspring of the best individual in all the evolved populations after 50,000 generations. The observed values are almost always just above that yielding an average of 1 neutral offspring with the selection

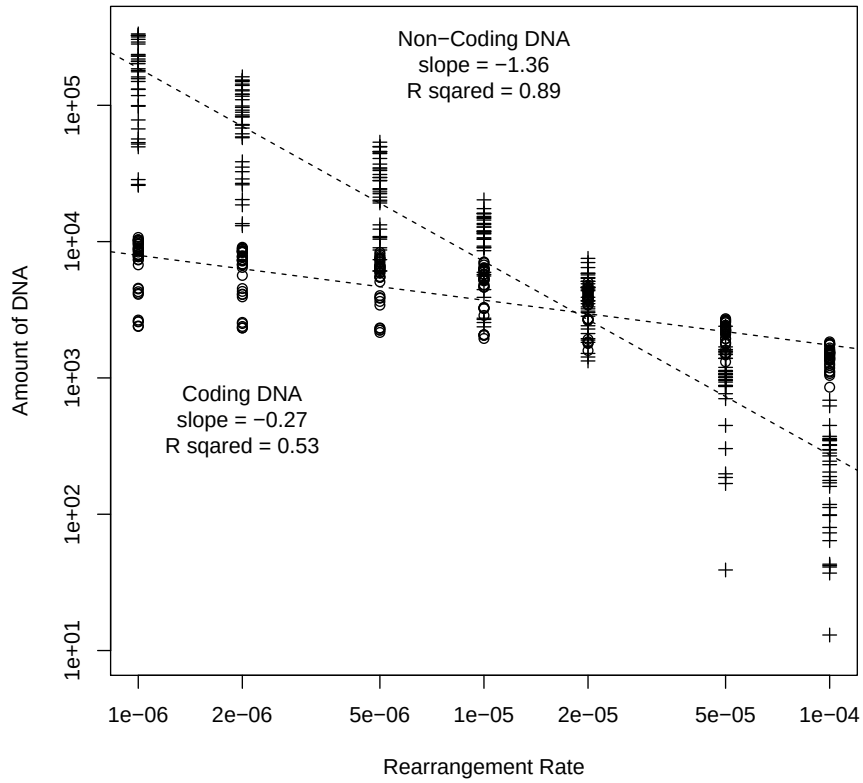
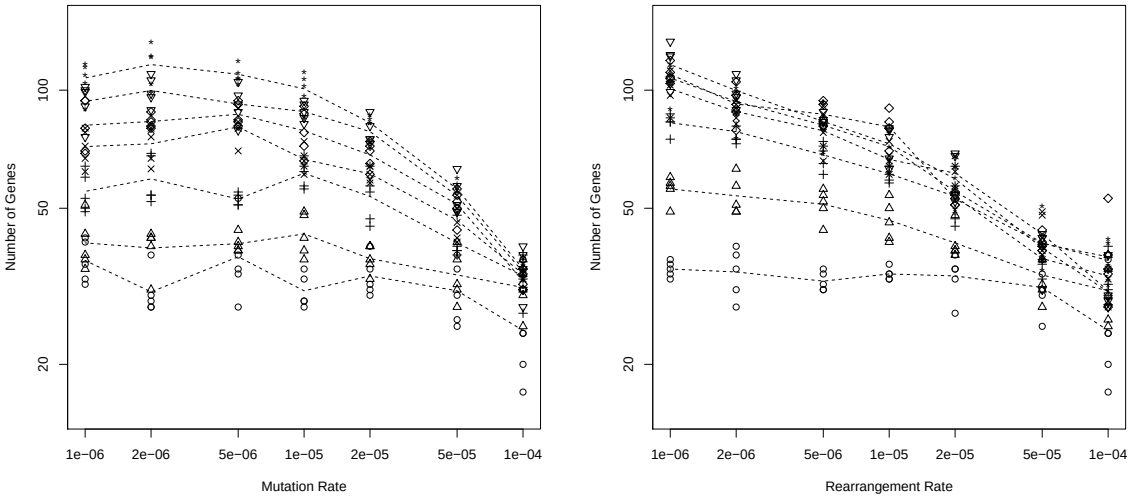


Figure II.4 – Amount of coding (circles) and non-coding (crosses) DNA of the final best organism of each simulation as a function of the rearrangement rate μ_r. The amount of both coding and non-coding sequences seems to follow roughly a power law (the variability for a given value of μ_r is at least partly due to the different values of μ_m). However, the slope for the coding sequences is very mild while that for the non-coding sequences is lower than -1, reflecting a super-linear relation.

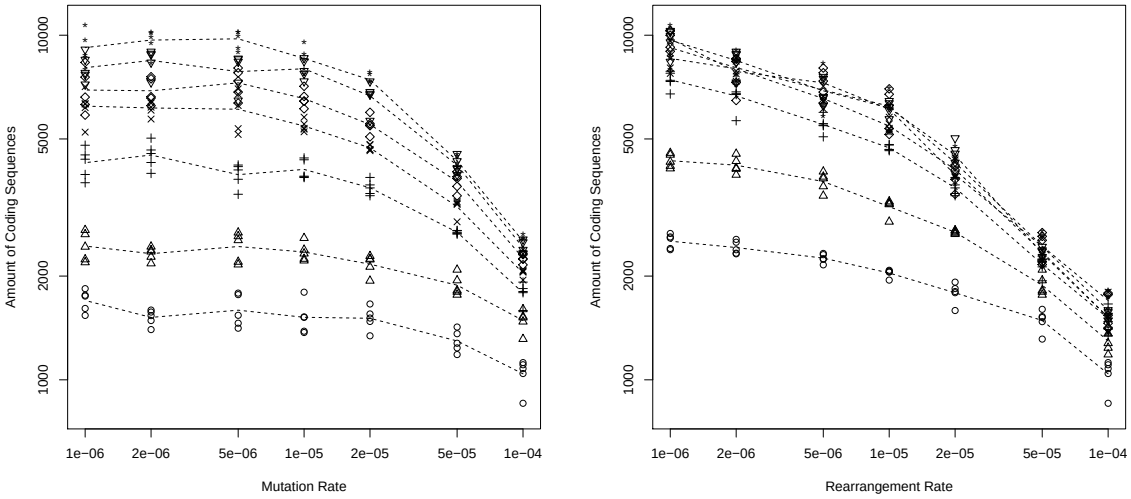
intensity that was used. This reflects the indirect selection of an appropriate trade-off between exploitation and exploration, the production of one neutral offspring ensuring that nothing is lost while maximizing the number of exploration trials. Note that the displayed data is that of the final best individual of each simulation, those organisms lying under this threshold could well suffer an error catastrophe and be replaced by more robust lineages. As a matter of fact, it would be of great interest to compare the evolutionary fate of lineages displaying different levels of F_ν.

3.3 Evolution of the Structure of the Transcripts

Considering more specifically transcription-related features, our attention was drawn by the surprising yet very clear trend for shorter genomes to contain longer cRNAs (figure II.8(a)). This trend is accompanied by a strong increase in the average number of coding



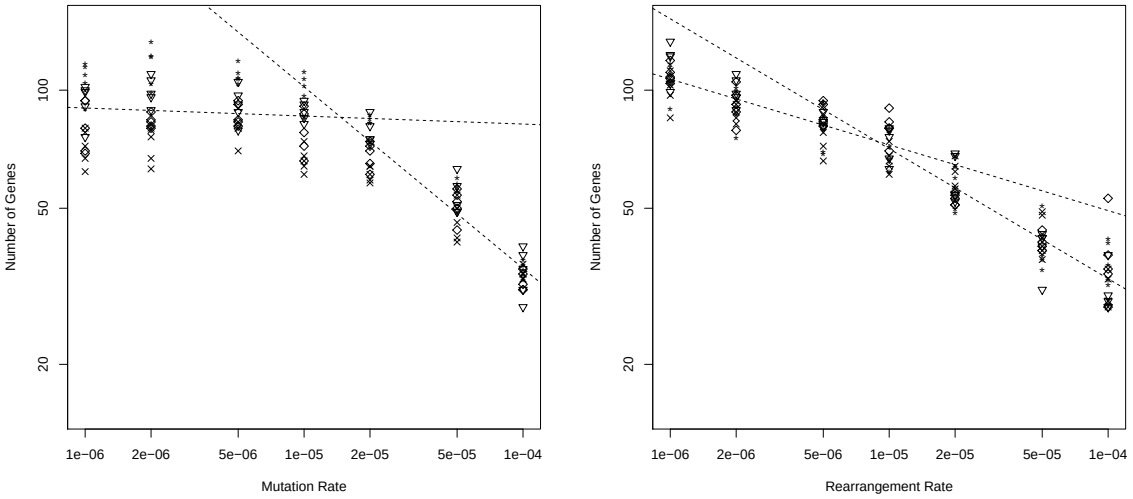
(a) Number of Genes



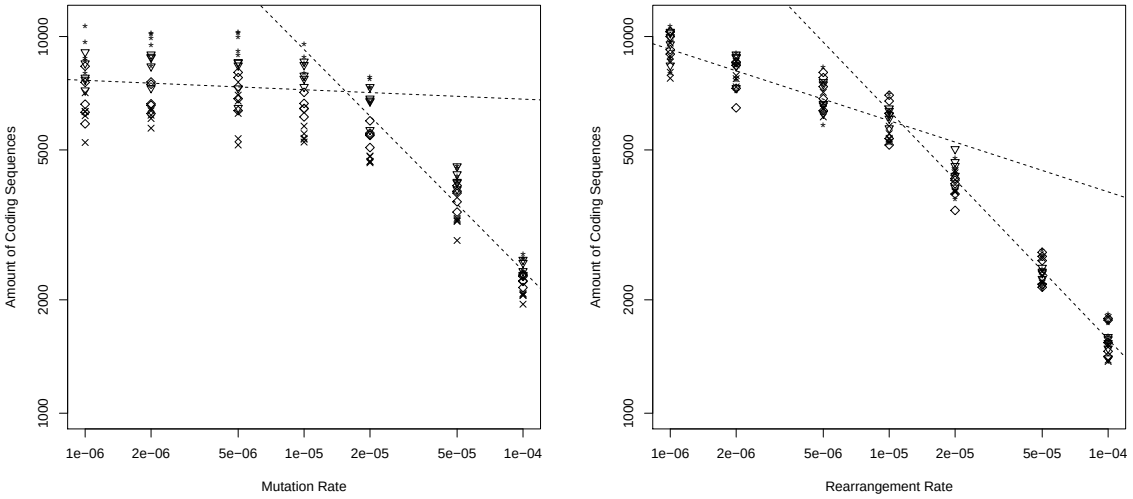
(b) Amount of coding DNA

Figure II.5 – Number of genes **(a)** and amount of coding DNA **(b)** of the final best organism of each simulation as a function of the mutation rate μ_m (left) and of the rearrangement rate μ_r (right). μ_r and μ_m are also reported as the shape of the points (circles: 10^{-4}, upwards pointing triangles: 5×10^{-5}, plus signs: 2×10^{-5}, multiply signs: 10^{-5}, diamonds: 5×10^{-6}, downwards pointing triangles: 2×10^{-6}, stars: 10^{-6}). μ_m and μ_r have a similar impact on coding sequences, as opposed to their very different effects on non-coding sequences (figure II.3).

sequences per coding RNA as is shown in figure II.8(b). Since the size of the genome scales as a power law with the rearrangement rate, both these effects are clearly dependent on the rearrangement rate.



(a) Number of Genes



(b) Amount of coding DNA

Figure II.6 – Number of genes **(a)** and amount of coding DNA **(b)** of the final best organism of each simulation as a function of the mutation rate μ_m (left) and of the rearrangement rate μ_r (right). μ_r and μ_m are also reported as the shape of the points (multiply signs: 10^{-5}, diamonds: 5×10^{-6}, upwards pointing triangles: 2×10^{-6}, stars: 10^{-6}). For reasons of clarity, we did not plot the points corresponding to the highest three values of μ_r on the left hand side figures and of μ_m and those on the right. It is clear that there is a threshold, both for μ_m and μ_r, above which both the number of genes and the amount of coding sequences fall more drastically when μ_m or μ_r are increased.

Since the average number of genes per cRNA increases with the rearrangement rate, it is not surprising to observe substantially fewer monocistronic RNAs at high rearrangement

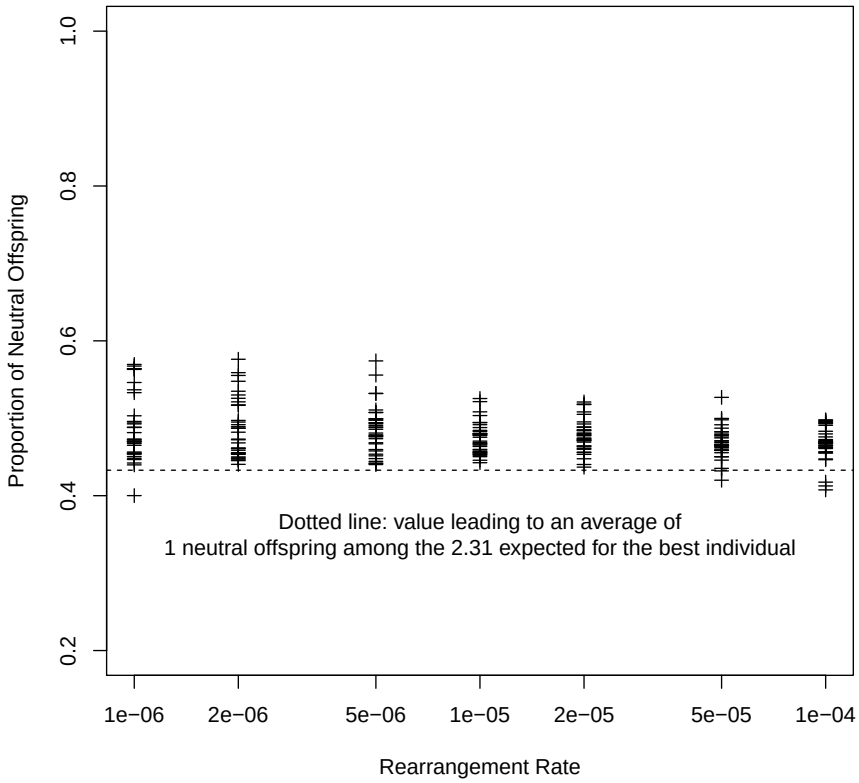
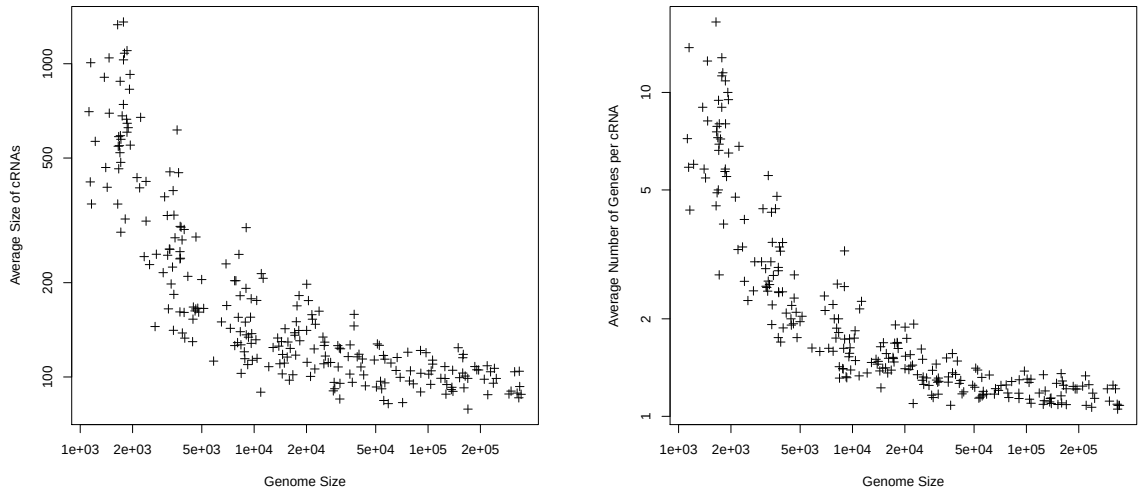


Figure II.7 – Proportion F_ν of neutral offspring of the final best organism of each simulation. Given the selection scheme (exponential ranking) and pressure (0.998), the best organism of any generation of any of these simulations will produce, on average, $W = 2.31$ offspring. The particular value $F_\nu = 0.43$ is remarkable since it leads to $F_\nu.W = 1$, *i.e.* it is the value with which the best individual of the population will produce one neutral offspring on average. Apart from a few exceptions, all the points are just above this particular value, which is represented by the dashed line on the figure. This data was obtained empirically by simulating 10,000 independent replications of the final best individual of each simulation.

rates than at low ones – figure II.9(a). We did not expect, however, that the number of polycistronic RNAs (operons) would remain stable throughout the whole set of tested parameters – figure II.9(b). In fact it is not so surprising that some simple operons arise as a result of pure chance, especially in genomes containing many genes, *i.e.* at low rearrangement rates, where operons only represent 10 to 20% of the coding RNAs (figure II.9(c)). Taken together, these observations suggest that, at high rearrangement rates, some mRNAs are greatly extended to include many genes, the former promoters transcribing these genes being subsequently removed, thus leading to a decrease in the number of monocistronic RNAs while greatly increasing the number of genes in at least some operons.

The dynamics that lead to this RNA lengthening during evolution are very interest-



(a) Average size of cRNAs

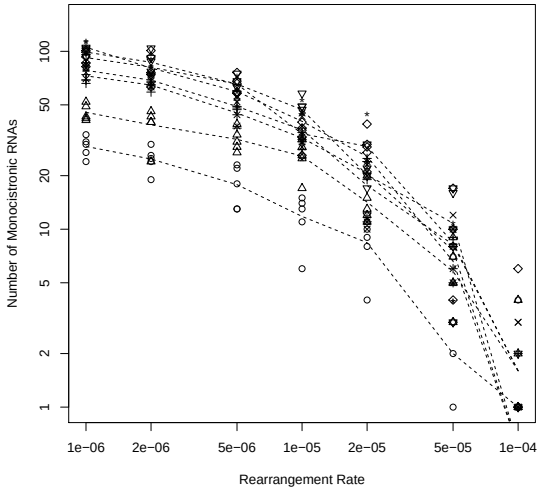
(b) Average number of genes per cRNAs

Figure II.8 – **(a)**: average size of coding RNAs and **(b)**: number of genes per coding RNA as a function of genome size. There is a clear tendency for very short genomes to have long coding RNAs containing many genes.

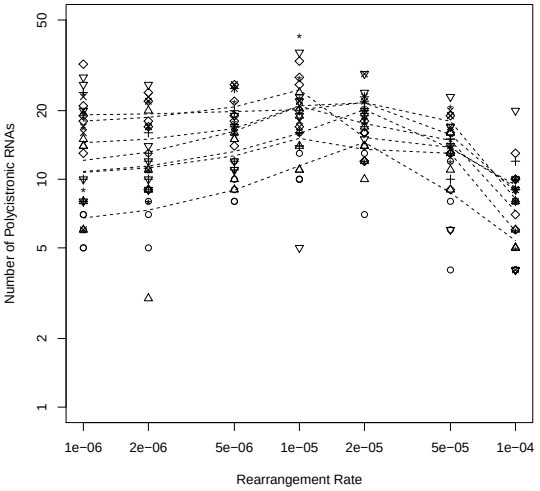
ing: figure II.10 shows that the higher the rearrangement rate, the more terminators are counter-selected. This was clearly expected since getting rid of terminators is the only way of producing longer RNAs. The effect on promoters however, is very surprising: up to a certain rate of rearrangements, promoters seem to be all the more selected for, as the rearrangement rate is high. Yet, above a critical threshold (around $\mu_r = 5 \times 10^{-5}$), the observed density of promoters shrinks drastically to values well under that expected for random sequences, thus underlying a clear counter selection of promoters at high rearrangement rates (figure II.10).
Finally, as figure II.11 shows, the number and proportion of non-coding RNAs are also strongly determined by the rearrangement rate. Actually, when simulations are grouped by mutation rates, the number of ncRNAs scales as a power law of the rearrangement rate – figure II.11(a). At one extreme, when μ_r is very low, over nine RNAs out of ten do not contain a single gene, while at the other extreme (high μ_r), genomes contain only a few ncRNAs, and in some cases, none at all.

4 Discussion

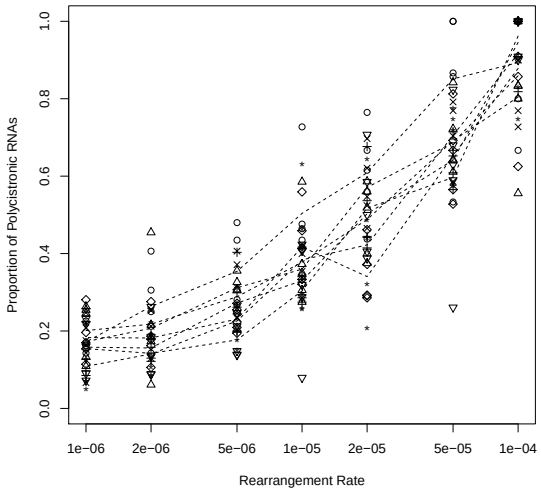
In the experiments we have presented here, 50,000 generations of evolution have produced very diverse organisms. Moreover, the different genomes that evolved reproduce the whole range of genome organizations observed in real organisms. At one extreme, organisms having evolved under high rearrangement rates present prokaryote-like genomes: very short and dense genomes bearing only a few genes usually transcribed through even fewer mRNAs and having almost no non-coding RNAs. At the other extreme, under low



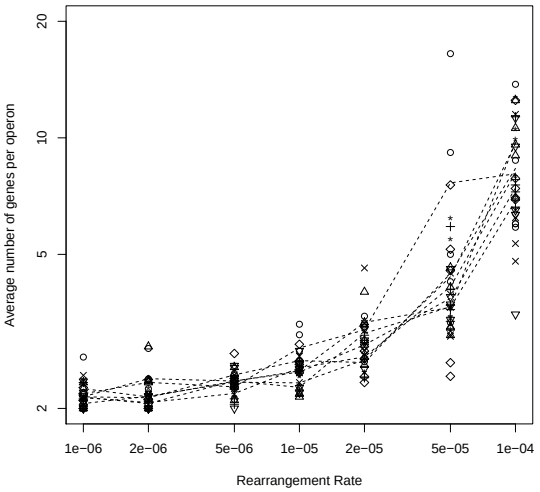
(a) Number of Monocistronic RNAs



(b) Number of Polycistronic RNAs

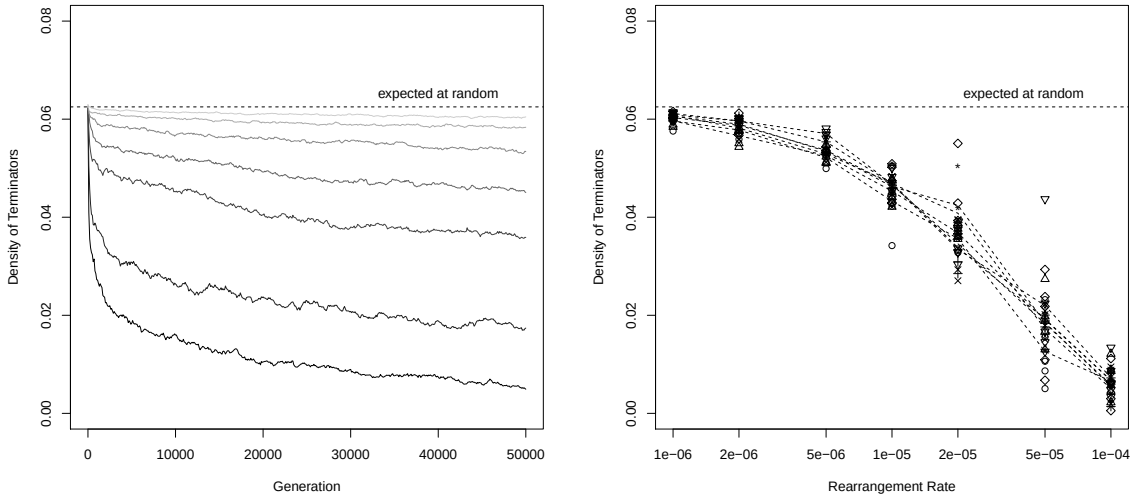


(c) Proportion of Polycistronic RNAs

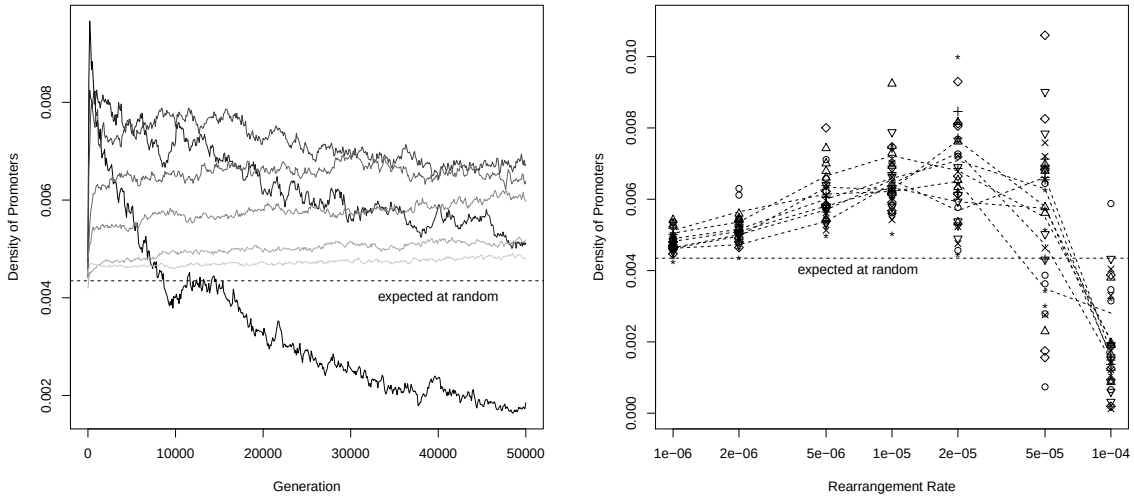


(d) Average number of genes per operon

Figure II.9 – Number of **(a)**: monocistronic and **(b)**: polycistronic RNAs as a function of μ_r. **(c)**: proportion of operons among cRNAs and **(d)**: average number of genes per operon as a function of μ_r. Different values for μ_m are reported as the shape of the points (circles: 10^{-4}, upwards pointing triangles: 5×10^{-5}, plus signs: 2×10^{-5}, multiply signs: 10^{-5}, diamonds: 5×10^{-6}, downwards pointing triangles: 2×10^{-6}, stars: 10^{-6}). Low values of μ_r lead to genomes containing mainly monocistronic RNAs and a few simple operons. High rearrangement rates lead to genomes having almost no monocistronic RNAs and a number of operons comparable to that obtained at low rates. However, operons that evolved at high μ_r are a lot more complex, containing way more genes than those evolved at low μ_r.

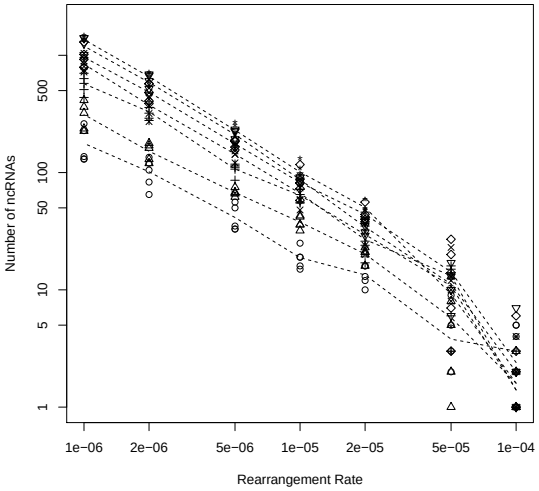


(a) Density of Terminators

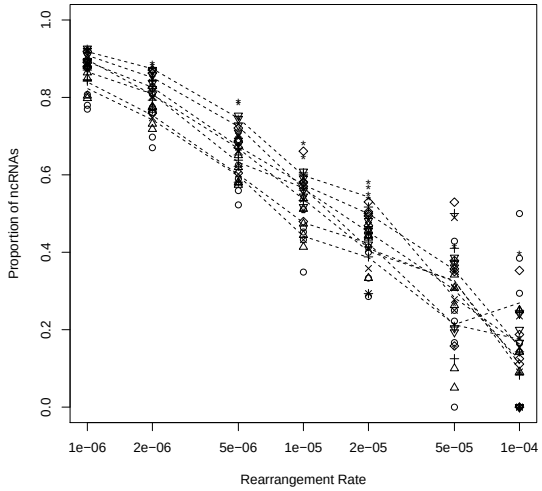


(b) Density of Promoters

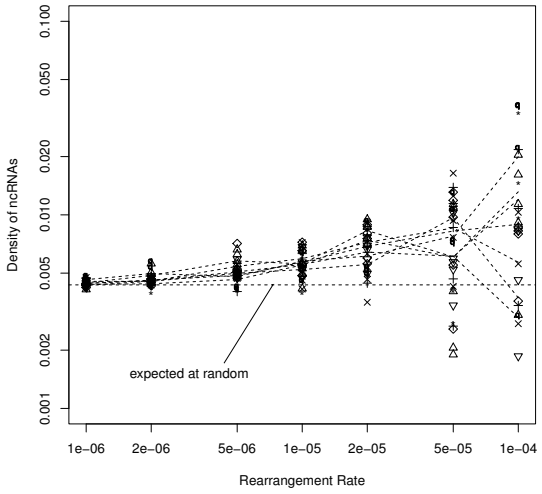
Figure II.10 – **(left)**: evolution of the density of terminators **(a)** and of promoters **(b)** in the lineage of the final best individual of each simulation. **(right)**: stabilized density of terminators **(a)** and of promoters **(b)** computed as the mean value for last 10,000 ancestors of the final best individual of each simulation. Different values for μ_m are reported as the shape of the points (circles: 10^{-4}, upwards pointing triangles: 5×10^{-5}, plus signs: 2×10^{-5}, multiply signs: 10^{-5}, diamonds: 5×10^{-6}, downwards pointing triangles: 2×10^{-6}, stars: 10^{-6}). While terminators are all the more counter-selected that μ_r is high, there seems to be a threshold in μ_r under which promoters are favoured and above which they are selected against.



(a) Number of ncRNAs



(b) Proportion of ncRNAs



(c) Proportion of ncRNAs

Figure II.11 – **(a)**: number and **(b)**: proportion of ncRNAs as a function of the rearrangement rate. **(c)**: density of ncRNAs in non-coding sequences. ncRNAs are very numerous at low μ_r while nearly absent at high μ_r. The density of ncRNAs in non-coding sequences seems to be close to the density expected for a random sequence (the greater diversity at high μ_r) can be explained by the small overall size of non-coding sequences in these cases).

rearrangement rates, evolution tends to favour huge genomes with many mRNAs, both coding and non-coding, the latter kind usually bearing one single gene. These genomes also contain a huge proportion of excess DNA.

Mutations, Rearrangements and Genome Structure

On the level of the genome, we have observed that both the local mutation rate and the chromosomal rearrangement rate have a clear impact on the organization of the genome. However, if it is mainly the rearrangement rate that determines the amount of non-coding sequences (excess DNA), coding sequences seem to be governed by both the mutation and the rearrangement rates. These observations point us to the *error threshold* principle and the *mutational burden* hypothesis. It clearly appears that the local mutation rate imposes an upper bound to the overall amount of coding sequences, regardless of non-coding sequences. An organism with a wider mutational target would undergo too many mutations within its coding sequences, likely producing no neutral offspring whatsoever. Then, because of the predominance of deleterious over beneficial mutations, all of its offspring would probably have lost fitness. As we showed in the previous chapter, because large duplications and large deletions involve DNA sequences whose mean size is proportional to the size of the genome, non-coding sequences are mutagenic for the genes they surround (Knibbe et al., 2007a). The rate at which these mutational events occur will then produce an error threshold-like effect: the rearrangement rate being fixed, a too large genome would undergo too many duplications and deletions, leading to only less-fit offspring, similarly to what happens in coding sequences. Thus, the rate at which chromosomal rearrangements occur constrains the overall amount of DNA (including non-coding sequences) under a certain threshold. Because the local mutation operators have almost no effect when they occur in non-coding sequences (the probability that they might create a new coding sequence *de novo* is very low), the local mutation rate has no effect on the size of the non-coding sequences. Large duplications and deletions however, have the potential of affecting the phenotype of an organism regardless of whether their breakpoints fall in coding or non-coding sequences, they hence affect both coding and non-coding sequences. All other mutation events have effects that are limited to coding sequences. As a conclusion, we could propose to consider local mutations as cis-acting genetic operators and chromosomal rearrangements as both cis- and trans-acting operators.

Evolution of Operons

On the level of the organization of the transcriptome, we observed the emergence of operon structures in every single simulation of this experiment. Since no horizontal transfer was allowed in these experiments, the selfish operons theory can easily be discarded as an explanation for the emergence of these operons. Indeed, horizontal transfer is a central and necessary condition for the emergence of selfish operons. One of the remaining candidates to account for the emergence of the observed operons is the co-regulation model, according to which genomes should be more modular than expected at random, operons containing preferentially functionally related genes. As for now, we did not test this hypothesis on this specific dataset. However, in the dataset published in (Parsons et al., 2010b), we tested this hypothesis by conducting a systematic pairwise comparison of the proportion of functionally related genes both within operons and on the whole genome(two genes were considered functionally related when the subset of biological processes they contributed to overlapped in Ω). Given that these experiments were conducted in a stable environment, no regulation was needed whatsoever. Yet, the results showed a moderate tendency of

functionally related genes to be packed together on the same operon: the proportion of pairs of functionally related genes was 1.26 times higher (median value) than the same proportion on the whole genome. Although the effect was small, the ratio was significantly different from 1 (non parametric sign test, $p - value = 7 \times 10^{-4}$). These results did not allow us to conclude either in favour of or against the co-regulation theory, we hence look forward to testing this hypothesis with the dataset we have presented here.

Figure II.9(c) showed us that the proportion of polycistronic mRNAs varied greatly according to the rearrangement rate the organisms were subjected to. Furthermore, the operons that evolved under high rearrangement rates tend to contain many more genes than those having evolved under low rates. This is relevant when considered in the light of the mutational burden hypothesis: as we have previously stated, the selection for a specific level of mutational robustness strongly constrains the size of the genomes. At high rearrangement rates, long genomes lead to error catastrophe which yields a long term selection of shorter genomes. On the other hand, the direct selection of the fittest organisms applies a strong pressure in favour of organisms having many genes, which allow them to better approximate the environmental function. Taken together, these two pressures result in the emergence of a composed pressure on the gene-density of the genomes. Now, there are different ways through which a genome can be densified. At moderate rearrangement rates, the optimal gene density can be achieved by simply acting on excess DNA, the coding sequences remaining mostly untouched. When the rates are very high however, the amount of excess DNA shrinks to almost nothing. A further compaction of the genome can then be achieved for instance by making genes overlap (either on the same strand or on both strands), which we indeed observed in most of the genomes that evolved in the context of high rearrangement rates. But our results also suggest that evolution found another way of further densifying the genome: getting rid of some of the transcription signals (promoters and terminators) and sharing the remaining signals between several genes.

Terminators in particular fragment the genome, forbidding the sequences directly downstream from them (on both strands) from being translated, until the next promoter. Each terminator on the genome hence unmistakably leads to a loss of gene density. Deleting a terminator that marked the end of a transcribed region will automatically cause this transcribed region to be extended to the sequence downstream from it, until the next terminator.

As for promoters, since they are relatively long (22 bp), mutualizing a promoter between several genes would allow to spare some space on the sequence compared to the case where each gene has its own promoter.

Let us consider two transcribed regions $TR1$ and $TR2$ defined by the promoters $prom1$ and $prom2$ and the terminators $term1$ and $term2$. Suppose that $TR1$ and $TR2$ are close to each other on the same strand, each bearing one gene, respectively $G1$ and $G2$. The destruction of $term1$ by any kind of mutation (for instance a point mutation or a small deletion) would cause $TR1$ to be extended to the whole region between $prom1$ and $term2$, that includes both $G1$ and $G2$, thus leading to the creation of an operon. $G2$ will then be transcribed by both $prom1$ and $prom2$ which means that destroying $prom2$ will not cause the loss of $G2$. Now in the specific case when $prom1$ and $prom2$ had the same transcription rate, destroying both $term1$ and $prom2$ during the same replication would have absolutely no effect on the phenotype, while creating an operon and deleting two

monocistronic RNAs. Interestingly, the simultaneous deletion of $term1$ and $prom2$ can easily be achieved by a single mutation, either a long deletion or even, if $term1$ and $prom2$ overlapped, by a point mutation or an indel.

According to this series of events, the dynamics of promoters and terminators should be similar, but a more common case is that only $term1$ is gotten rid of, which produces $G2$ to be translated twice, thus rising its expression level. This more simple event leads to the transformation of a monocistronic RNA into an operon and can participate in the counter-selection of terminators while having no impact on promoters.

Note that the opposite case (the deletion of only $prom2$, $term1$ remaining in place, would cause the loss of $G2$ (that would become a pseudo-gene). Since the reversion of this mutation is highly improbable, $G2$ is very likely to be irrevocably lost, which in turn is very likely to be deleterious. There is however a case where the loss of $G2$ could become fixed in the population: this would happen if the direct cost of losing this gene were counter-balanced by the increase in robustness it may have caused. This is likely to occur at high rearrangement rates where there is a very strong pressure towards genome compaction. This overcoming of indirect selection could well be the explanation of the counter-selection of promoters at high rearrangement rates. We indeed observe that the organisms having evolved at high rearrangement rates are less well-adapted yet more robust than those having evolved with low rates, suggesting this kind of effects. Still, asserting this hypothesis would require further experiments and analyses to be performed.

Non-Coding RNAs

Although a very small proportion of eukaryotic genomes is translated into proteins, a substantial fraction of these genomes is nonetheless transcribed, mostly into non-coding RNAs. Not all of these ncRNAs have a known function and a great deal of effort is put into identifying these putative functions. In our model, we have an absolute control on the functions RNA *can* have. Specifically, in our model, ncRNAs have absolutely no function. Yet, while they are seldom found on short and dense genomes, they are very common when rearrangement rates are low, *i.e.* on large, mostly non coding genomes. Interestingly, they are found at a proportion close to that which would be expected in a random sequence. It hence seems that ncRNAs are naturally present in intergenic regions, constituting a constantly available pool of substance that we know can acquire new functions in real organisms (*e.g.* the post-transcriptional regulatory activity of micro RNAs). It is also tempting to suggest that because ncRNAs are naturally present in any DNA sequence, they constitute a good substrate for the appearance of novel genes, and hence that they could be selected for in the long term because they promote evolvability. However this question will require a precise analysis of the dynamics of gene acquisition for which we will need to develop specific tools. As for now, we can only say that they are not selected for since their proportion in large genomes is not greater than that expected in random sequences.

5 Conclusion and Perspectives

In this chapter, we have presented results that clearly reproduce features of genome organization as it is observed in real organisms, in particular the emergence of operon structures that seems to be favoured in the context of high rates of chromosomal rearrangements. The emergence of these operons specifically under high rearrangement rates points us to the mutational burden hypothesis, where a second-order selective pressure for a specific level of mutational robustness leads to genome streamlining.
We now plan to analyse the modularity of the genomes in these experiments and to conduct further experiments using an extension of the model that includes explicit regulation of gene expression (R-Aevol, see next chapter) to determine to what extent the co-regulation model can participate in the creation and maintenance of operon structures. We also plan to conduct experiments allowing for horizontal transfer in order to test the selfish operon hypothesis. Finally, we are looking forward to developing new tools to conduct detailed analysis of the precise mutations that went to fixation. This would allow us to study in detail the dynamics of gene acquisition and, in certain conditions, gene loss, that are of major importance to better understand both the dynamics of operon formation and the putative role of ncRNAs as innovation hotspots.

Chapter III

Indirect Selection and the Regulation of Gene Expression

The results presented in this chapter have been published in Beslon et al. (2010b) and Beslon et al. (2010a)

1 Introduction

Evolution provides living organisms with a way of adapting to their environment in the long term. On the scale of an organism's lifetime, however, evolution is of little help. During its lifetime, an organism may be confronted with different environments. An organism able to react to environmental changes would then be better-adapted than one that isn't. Evolution's answer to that takes many forms. One can of course think of one's sensing organs as a patent example. However, bacteria usually lack a brain to analyze such signals. A more universal answer to this need for a response to external changes is the regulation of gene expression.

Genetic regulation networks are recognized to be one of the main control centres in cells and organisms. They are also well known to be very complex and intricate throughout many different species. The conjunction of these two factors made the understanding of these networks a very active field of research in the last few decades. A great deal of work has shown that the structure of these regulation networks is far from random: regularities were found at all scales, from small motifs (Alon, 2007; François and Hakim, 2004) to global connectivity patterns (Barabási and Oltvai, 2004; Zhu et al., 2007). Understanding these regularities and deciphering the emergent dynamics of these networks are among the most fascinating challenges of modern biology and have been central questions of systems biology ever since this research field emerged ten years ago.

Being a product of evolution, gene regulation networks have but little to do with engineered networks: one can hardly identify independent modules that would undertake

clearly defined tasks in the system. Rather, regulation networks were built by a process of trial and error, yielding very intricate structures within which several functionalities are often superimposed in overlapping sub-networks. Systems biology is often considered a reverse engineering process applied to biological entities: observations are made regarding the behaviour of the studied system in different external conditions (and possibly, its response to man-made perturbations – Ideker et al. 2006) in an attempt to uncover the organizational principles of the system. However, when reverse engineering can usually assume the system to have been conceived following reasonable conception rules and reasoning, it is not true when it comes to biological systems. Yet, evolution has its own rules: it was shown, for instance, that under some specific conditions (*e.g.* cyclic environments), evolution *can* produce an organized system, similar to what an engineer might come up with (Alon, 2003; Kashtan and Alon, 2005). The existence of other general laws that could govern the organization of biological networks depending on external conditions is an open question. Our grand challenge is hence to identify the "language" that evolution has created for regulation networks and how it can be translated from a structural description (*i.e.* the set of weighted links, motifs and modules) to a functional description (cell behaviour – Wolf 2003).

These questions are very difficult to tackle with real organisms, either because they require long and complex experimental setups or because results are difficult to analyze given the little knowledge available. As far as regulation networks are concerned, computational evolution has been used to investigate the evolvability of networks (Crombach and Hogeweg, 2008) or the development of modular structures under cyclic environmental conditions (Kashtan and Alon, 2005). One of the best-known models, namely the GRN model, was proposed by Wolfgang Banzhaf (Banzhaf, 2003) and used to investigate the emergence of specific topological properties in regulatory networks (Kuo et al., 2006). More recently, Claudio Mattiussi and Dario Floreano proposed the "Analog Genetic Encoding" framework (Mattiussi and Floreano, 2007) which was later on used to investigate the modular structure of regulation networks (Marbach et al., 2009). Other authors have used computational evolution in order to evolve small networks performing predefined tasks as oscillators (Knabe et al., 2008) or switches (François and Hakim, 2004).

Here, we are particularly interested in exploring the putative effects of indirect selective pressures on gene regulation networks. As we have previously stated, a strong indirect selective pressure was identified using the Aevol model, that constrains the size and structure of the genome to achieve a specific level of mutational variability of the phenotype. Whether such a pressure can still be at play in a system in which an additional degree of freedom is provided by a regulation process, and its implications on the network itself are fascinating questions. To tackle these questions, the Aevol model was extended to include an explicit process of regulation at the level of transcription. This extended model, called R-Aevol, was mainly developed by Yolanda Sanchez-Dehesa during her PhD thesis (Sanchez-Dehesa, 2009). We provide here a description of the R-Aevol model followed by two sets of experiments conducted respectively with a trivial and a demanding environment. The results not only confirm the second-order selective pressure that had been previously observed with the Aevol model, but also show that this pressure can also drive the size and complexity of regulation networks.

2 Introducing Regulation in Aevol: the R-Aevol model

In real organisms, the regulation of gene expression can be achieved through several different mechanisms acting at different stages of gene expression. The most famous and best understood of these mechanisms was discovered by Jacob and Monod in 1960 along with the operon structure (Jacob et al., 1960). The transcription of DNA into RNA can be up- or down-regulated when some particular proteins called Transcription Factors (TFs) bind to the DNA at specific sites upstream or downstream from the promoter, making it either easier or more difficult for the RNA-Polymerases to proceed to the transcription. Regulation sites upstream from the promoter can either facilitate or impede the initiation of transcription while regulation sites downstream from the promoter will preclude the elongation of the RNA being synthesized. Now, in total absence of transcription factors (*i.e.* when there is no regulation), the transcriptional activity of a prokaryotic promoter is at its *basal level* or *ground level* (Struhl, 1999). The regulation of the transcriptional activity of RNA-Polymerases, either up or down from the ground level, results in a variation in the concentration of the corresponding RNA and subsequently in the rate of translation of the genes it carries, finally yielding a change in the encoded protein's concentration.

Whether the expression of a gene is up- or down-regulated depends on which site the transcription factor is bound to and on the transcription factor itself. A transcription factor that binds to the operator site (downstream from the promoter) will prevent the RNA from being elongated, *i.e.* it will forbid the transcription to be completed, leading to a down-regulation of the genes carried by this RNA. TFs that bind to a regulation site upstream from the promoter will impact transcription initiation by either facilitating or repressing the binding of RNA-Polymerases. TFs interacting with an upstream site can then induce either an increase or a decrease in the expression of the corresponding genes.

R-Aevol is an extension of Aevol that includes a model of prokaryotic regulation in the artificial chemistry. To model the interactions between transcription factors and promoters in R-Aevol, we defined two binding sites for each promoter: the 20 base-pair long sequences directly flanking the promoter. Preceding the promoter, the *enhancer* site will allow TFs that bind to it to increase the transcriptional activity. The *operator* site, directly following the promoter, will on the contrary produce a down-regulation of the promoter's activity whenever a TF binds to it. In R-Aevol, whether a given protein is able to bind to a specific site is determined by a value of "affinity" between the Amino-Acid (AA) chain of the former and the genetic sequence of the latter. More precisely, a protein can bind to a given binding site if its sequence contains at least one *regulation domain* for this site. In R-Aevol, regulation domains are defined as small 5 AA long motifs, each AA of which, when aligned with the sequence of a binding site, has a strictly positive affinity value with the base quadruplet it faces (see figure III.1). These individual AA-to-base-quadruplet affinities are given by an "affinity matrix" B which is initialized once and for all at the beginning of the simulation. Since the affinity of a 5 AA motif with a regulation site is computed as the product of values taken in B, the probability of any such motif actually being a regulation domain of a given binding site will be determined by the proportion of null values in this matrix. We hence defined a parameter of the model that makes it possible to tune the proportion of values of B that will be forced to 0, the remaining values being randomly filled with values in $[0, 1]$).

The global affinity of a protein with a binding site is that of its best regulation domain with

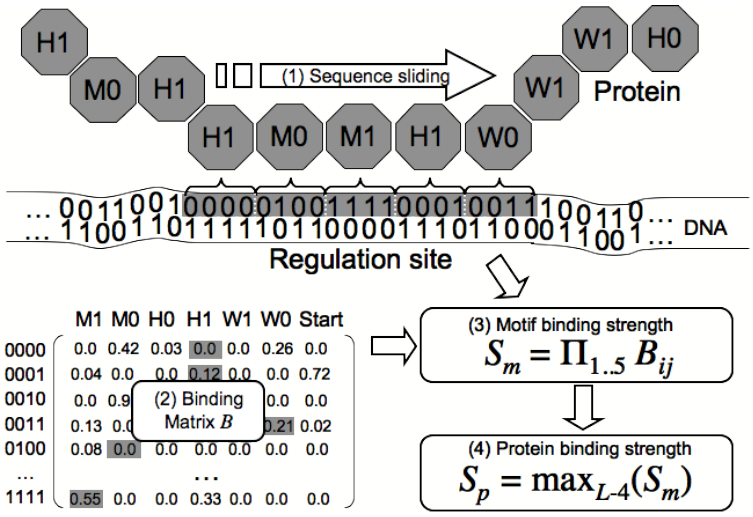


Figure III.1 – (from Beslon et al., 2010b). Computation of the affinity between TFs and regulation sites. **(1)**: the protein primary sequence slides in front of the 20-bp regulation site and all 5-AA-long motifs are tested. **(2)**: for each couple (AA, base quadruplet), the binding value B_{ij} is read in a binding matrix B (see main text for the initialization of the binding matrix). **(3)**: the binding strength, S_m, of the whole motif is the product of the five B_{ij} values and **(4)**: the binding strength, S_p, of the whole protein is the maximum strength over the $L-4$ motifs it contains (L being the length of the protein primary sequence). The regulatory activity of the protein p then depends on the global strength value S_p: for a given promoter i, a protein p is aligned both upstream and downstream from the promoter, resulting in two different S_p values (S_{p_up} and S_{p_down}). The upstream alignment enables us to compute the enhancing activity of the protein ($A_{pi} = S_{p\_up}$) while the downstream alignment gives its inhibition abilities ($I_{pi} = S_{p\_down}$).

regard to this site. In other words it is equal to the maximum affinity of all the possible 5 AA long motifs on the protein with the genetic sequence of the binding site. This value of affinity will determine the strength of the protein's influence on the transcriptional activity of the promoter it binds to, either increasing or decreasing it depending on whether it can bind to the enhancer site or the operator site (or sometimes to both).
Besides the ground level β_i (equation III.1) of a promoter, which depends on how close its sequence is to a consensus sequence [1], the transcriptional activity of a promoter depends on the combined activity of the transcription factors that activate it (equation III.2), and of those that inhibit it (equation III.3), A_{ji} (resp. I_{ji}) being the affinity of protein j with the enhancer site of the promoter i (resp. on its operator) and $c_j(t)$, the concentration of protein j at time t.

$$\beta_i = 1 - \frac{d_i}{d_{max} + 1} \tag{III.1}$$

with d_i, the Hamming distance between this particular promoter and the consensus se-

1. The ground level of a promoter in R-Aevol corresponds to the level of expression of a promoter in Aevol (see chapter I, section 5.1).

quence and d_{max}, the maximum Hamming distance allowed for the sequence to be a promoter.

$$A_i(t) = \sum_j c_j(t) A_{ji} \tag{III.2}$$

$$I_i(t) = \sum_j c_j(t) I_{ji} \tag{III.3}$$

The transcription rate e_i over time of an RNA is then given by the Hill-like function:

$$e_i(t) = \beta_i \cdot \left(\frac{\theta^n}{I_i(t)^n + \theta^n} \right) \cdot \left(1 + \left(\frac{1}{\beta_i} - 1 \right) \left(\frac{A_i(t)^n}{A_i(t)^n + \theta^n} \right) \right) \tag{III.4}$$

where n and θ are constant coefficients that determine the shape of the Hill-function.
Finally, given the transcription rate, one can compute the protein concentration (for the sake of simplicity, we assume here that the protein concentration is linearly proportional to the RNA concentration) through a synthesis-degradation rule (equation III.5). Thus, when a protein is regulated, its concentration is scaled up or down depending on its transcription rate.

$$\begin{cases} c_i(0) = \beta_i \\ \\ \frac{\partial c_i}{\partial t} = e_i(t) - \phi c_i(t) \end{cases} \tag{III.5}$$

where ϕ is a temporal scaling constant representing the protein degradation rate.
At each time step, the regulatory activity of each protein over each promoter is computed depending on the binding affinity between the protein and the promoter's regulation sites and on the concentration of the protein. Then, the concentrations are updated according to equation III.5 on the basis of a simple synchronous Euler integration scheme. In the current version of the model, the initial concentration of each protein is equal to the basal level of the corresponding promoter. However, we are currently developing a new version with inheritance, in which the initial protein concentrations of an organism are equal to the final concentrations of its parent.
Because protein concentrations vary over time, the phenotype of an organism is also defined dynamically as the possibility degree for each possible biological process at each time step t. As in the Aevol model, the environment in R-Aevol is represented as a phenotypic target. However, contrary to Aevol, organisms in R-Aevol "live" for a given time throughout which the environment may itself vary. The environment is thus represented similarly to the phenotype as the optimal possibility degree for each possible biological process at each time step. The fitness of an organism is then computed according to the combined *metabolic errors* of an organism with respect to the current state of the environment at different time steps. Moreover, external signalling molecules are introduced into the organism at specific time steps (typically when the environment is changed), thus modelling its sensing of the environment. These molecules are manually-designed amino-acid chains, *i.e.* proteins. These proteins may or may not have a metabolic function, but they must contain a regulation domain in order to be able to interact with the regulation network.

3 Gene Regulation Networks in a Trivial, Steady Environment

Eventually, our goal is to use R-Aevol to understand how regulation networks evolve depending on external conditions and on the complexity of the environment (*e.g.* number of states, frequency or periodicity of environmental variations). R-Aevol makes it possible to conduct evolution experiments in more or less demanding environments in which the organisms have to sense the possible changes in the environment through "molecular" signals. However, our first aim was to assess whether organisms evolving in simple steady environments would evolve basic regulation networks (even though they are not necessary in such constant environments). Another question we wanted to answer was whether the indirect selective pressure that was identified in Aevol would still be involved if organisms were able to regulate the expression of their genes, providing them with a new degree of freedom, and maybe an alternative path towards robustness or evolvability (Beslon et al., 2010b).
To test this hypothesis, we repeated in R-Aevol, the experiments conducted in Aevol (see chapter I, section 9.2 and chapter II). We left 18 independent populations of 1,000 individuals to evolve in a constant environment (the same environment we classically use for the experiments using Aevol – see figure I.2), the phenotype of each organism being computed during twenty time steps and its metabolic error corresponding to the mean gap during the last ten time steps. Six different sets of parameters were tested with three repetitions each, where the only changing parameter was the common mutation/rearrangement rate μ_{mr}, for which we tested the values 5×10^{-6}, 10^{-5}, 2×10^{-5}, 5×10^{-5}, 10^{-4} and 2×10^{-4} per base-pair per replication per type of mutation/rearrangement. Here, we could not isolate the local mutation rate from the chromosomal rearrangement rate as was the case in the experiments presented in the previous chapter. As R-Aevol requires a lot more computational time than Aevol, testing all the possible combinations of these parameters was impossible in a reasonable time. For the same reason, these experiments were conducted for only 15,000 generations. The values we used in these experiments for the main parameters are presented in table III.1.

3.1 Results

We analysed the structure of both the genomes and the regulation networks after $15,000$ generations. Again, we observed that many features of the evolved organisms are influenced by the mutation rate, the organisms having evolved under low rates of mutations and rearrangements being much more complex than those having evolved under higher rates. Figure III.2 shows the evolution of the metabolic error, the number of genes and the size of the genome of the best individual of each simulation while the genomes (with the genes they bear) of three representative organisms after 15,000 generations of evolution, under low, moderate and high mutation/rearrangement rates are presented in figure III.3. These results confirm those previously obtained using Aevol (without regulation – see (Knibbe et al., 2007a), chapter I, section 9.2 and chapter II): organisms having evolved under low rates of mutations and rearrangements have huge genomes containing many genes (93 genes on the genome shown in figure III.3(a)) while having a very high proportion of excess DNA (figure III.3(a): 97% of the genome). On the other hand, organisms having

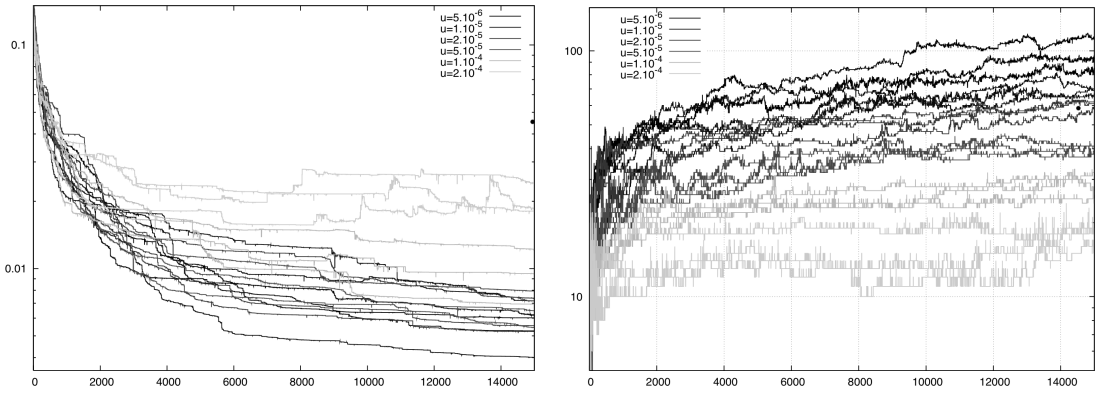
Parameter	Value
N	1,000
nb_gener	15,000
$init\_length$	5,000
$init\_method$	Clonal, One Good Gene
$selection\_scheme$	Exponential Ranking
c	0.995
$E = \sum_i \alpha_i G_i$	$\alpha_1 = 1.2; G_1 : \mu = 0.52; \sigma^2 = 0.12$
	$\alpha_2 = -1.4; G_2 : \mu = 0.2; \sigma^2 = 0.07$
	$\alpha_3 = 0.3; G_3 : \mu = 0.8; \sigma^2 = 0.03$
$env\_sampling$	300
μ_{point}	$\mu_{mr} \in \{5 \times 10^{-6}, 10^{-5}, 2 \times 10^{-5}, 5 \times 10^{-5}, 10^{-4}, 2 \times 10^{-4}\}$
μ_{s_ins}	
μ_{s_del}	
μ_{dupl}	
μ_{del}	
μ_{inv}	
μ_{trans}	
max_indel_size	6
W_{max}	0.033333333

Table III.1 – Parameters used in the experiments presented in this section. Mutation and rearrangement rates take their values among those proposed, one common value for each types of operators.

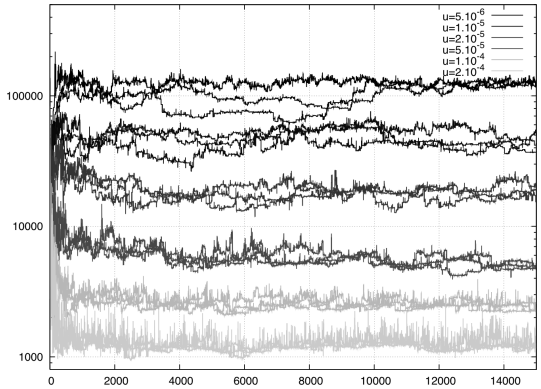
undergone very high rates of mutations and rearrangements have very small genomes containing only a few genes (38, resp. 16 genes on the genomes shown in figure III.3(b)) and almost no non-coding sequences (figure III.3(b): 65% and 37% of the genome for moderate and high μ_{mr} respectively).

In figure III.4, the amount of both coding and non-coding sequences of the best organisms after 15,000 generations are presented. Once more, as we had previously observed, the overall size of both the coding and non-coding sequences scale as a power law of the mutation/rearrangement rate, the slope for the coding sequences being quite mild while that for the non-coding sequences reflects a super-linear relation.

The analysis of the regulation networks is of great interest: in the experiments presented here, the environment is constant during the whole lifetime of the organisms, one could hence expect scarcely to observe any regulation networks at all. However, despite this lack of direct pressure for regulation, many evolved organisms present very large and complex regulation networks. In fact, we observed that the size and complexity of the networks were clearly correlated with the mutation/rearrangement rate, high values of μ_{mr} leading to very small and scarcely connected networks while low values of μ_{mr} produce large and complex networks. Figures III.5 and III.6 show examples of networks having evolved under low, moderate and high mutation rates. These networks correspond to the genomes presented in figure III.3.



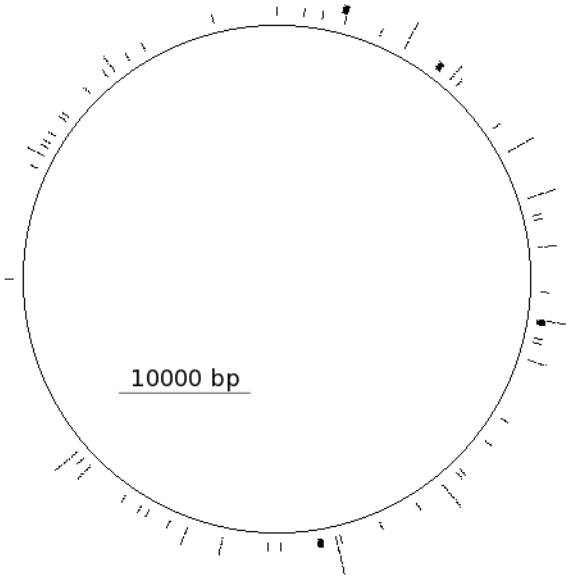
(a) Evolution of the metabolic error (b) Evolution of the number of genes



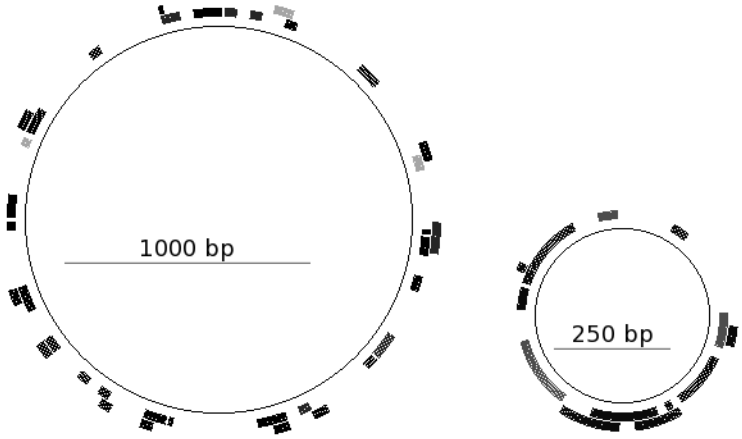
(c) Evolution of the size of the genome

Figure III.2 – (from Beslon et al., 2010b). Evolution of **(a)**: the metabolic error, **(b)**: the number of genes and **(c)**: the size of the genome of the best organisms during 15,000 generations. Grey scales represent the mutation/rearrangement rates, light grey corresponds to $\mu_{mr} = 2 \times 10^{-4}$ and black, to $\mu_{mr} = 5 \times 10^{-6}$. Again, all these measures depend on μ_{mr}. Note that even after only 15,000 generations, all the simulations seem to have stabilized.

The total number of genes being negatively correlated with μ_{mr} (as it is the case in experiments conducted with the Aevol model), it is not surprising that the size of the networks follows the same kind of relation. However, it is interesting to note that the number of transcription factors (TFs – genes whose products regulate the transcriptional activity, at least at one promoter site) increases faster than the total number of genes when mutation rates are lowered – figure III.7(a). In the smallest network – figure III.6, right –, only two genes out of 16 are TFs, corresponding to only 12.5% of genes. In the network having evolved under a moderate μ_{mr}, 16 genes out of 38 are TFs (42%) while in the largest network – figure III.6 –, 64 genes out of 93 are TFs (69%). In other words, there is a greater proportion of TFs when the mutation rate is low than when it's high. As shown in figure III.7(a), both the number of genes and of TFs seem to scale as power-laws of the mutation/rearrangement rate, the slope for TFs being steepest. This trend is even more clear if we consider only "pure TFs" (proteins that have a regulatory activity but no metabolic contribution – figure III.7(b)).



(a) A low mutation rate ($\mu_{mr} = 5 \times 10^{-6}$) leads to large genomes (here 120583 bp) with huge non-coding regions (here 97% of the genome).



(b) A medium mutation rate (left, $\mu_{mr} = 5 \times 10^{-5}$) leads to medium size genomes (here 4964 bp) with large non-coding regions (here 65% of the genome). A high mutation rate (right, $\mu_{mr} = 2 \times 10^{-4}$) leads to smaller genomes (1180 bp) with smaller non-coding regions (37%).

Figure III.3 – (from Beslon et al., 2010b). After 15,000 generations, the genomes range from large ones **(a)** to intermediate and small ones **(b)** depending on the mutation/rearrangement rate μ_{mr}. These differences are due to robustness and evolvability constraints: large genomes are not maintained when organisms face high rearrangement rates. On the contrary, under low rates, large genomes are more evolvable (see (Knibbe et al., 2007a) and Discussion). On each figure the circle represents the whole genome (scale is different on each figure). Grey arcs represent the coding regions (the grey level code is arbitrary, similar grey levels representing similar metabolic functions (*i.e.* proximity in the Ω space – see Chapter I, section 4).

3.2 Discussion

The way regulation was modelled in R-Aevol allows for a direct comparison between the structures that evolved in the model and similar structures in prokaryotes. Now,

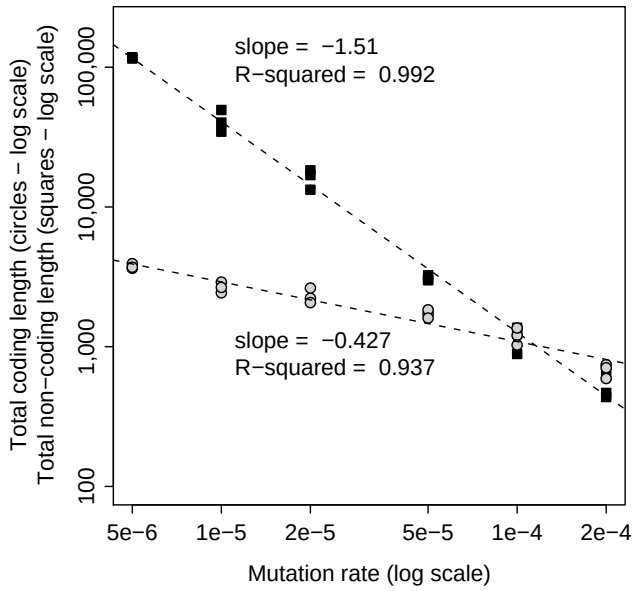


Figure III.4 – (from Beslon et al., 2010b). Amount of coding (grey circles) non-coding (black squares) sequences for the best organisms of the 18 simulations at generation 15,000 (log-log plot). Both values clearly scale with the mutation rate.

prokaryote genomic structures can be very diverse, with genome sizes ranging from ∼500 kb for the endosymbiont *Buchnera aphidicola* (Viñuelas et al., 2007) to more than 6 Mb for *Pseudomonas aeruginosa* (Stover et al., 2000). Similarly, the number of genes ranges from a few hundred (∼600 for *B. aphidicola*) to more than 5500 for *P. aeruginosa*. Variations in the functional content of the genomes are also visible at the transcription level: some organisms (e.g. *B. aphidicola*) are hardly able to regulate their transcriptional activity (Reymond et al., 2006) while others display complex regulation networks made up of thousands of tightly interconnected nodes (Stover et al., 2000). Comparative analyses of bacterial genomes show the diversity of genomic structures in an even more striking way. Through the analysis of annotated sequences, it was shown that the number of genes of different functional categories (genes involved in metabolic, regulation or transcription-translation processes) scale as power-laws of the total number of genes in the genome and that the exponents of these laws depend on the functional role of the family: the number of transcription factors (TFs), in particular, scales quadratically with the total number of genes while metabolic genes scale at most linearly with it – see (van Nimwegen, 2003; Molina and van Nimwegen, 2008) and figure III.8. Moreover, this increase is also correlated to the size of the genome (Konstantinidis and Tiedje, 2004). Note that, in the model, there are no genes involved in translational activities since these functions are coded in the core of the model. These results suggest that the intricacy of regulation networks grows faster than the size of the network itself.

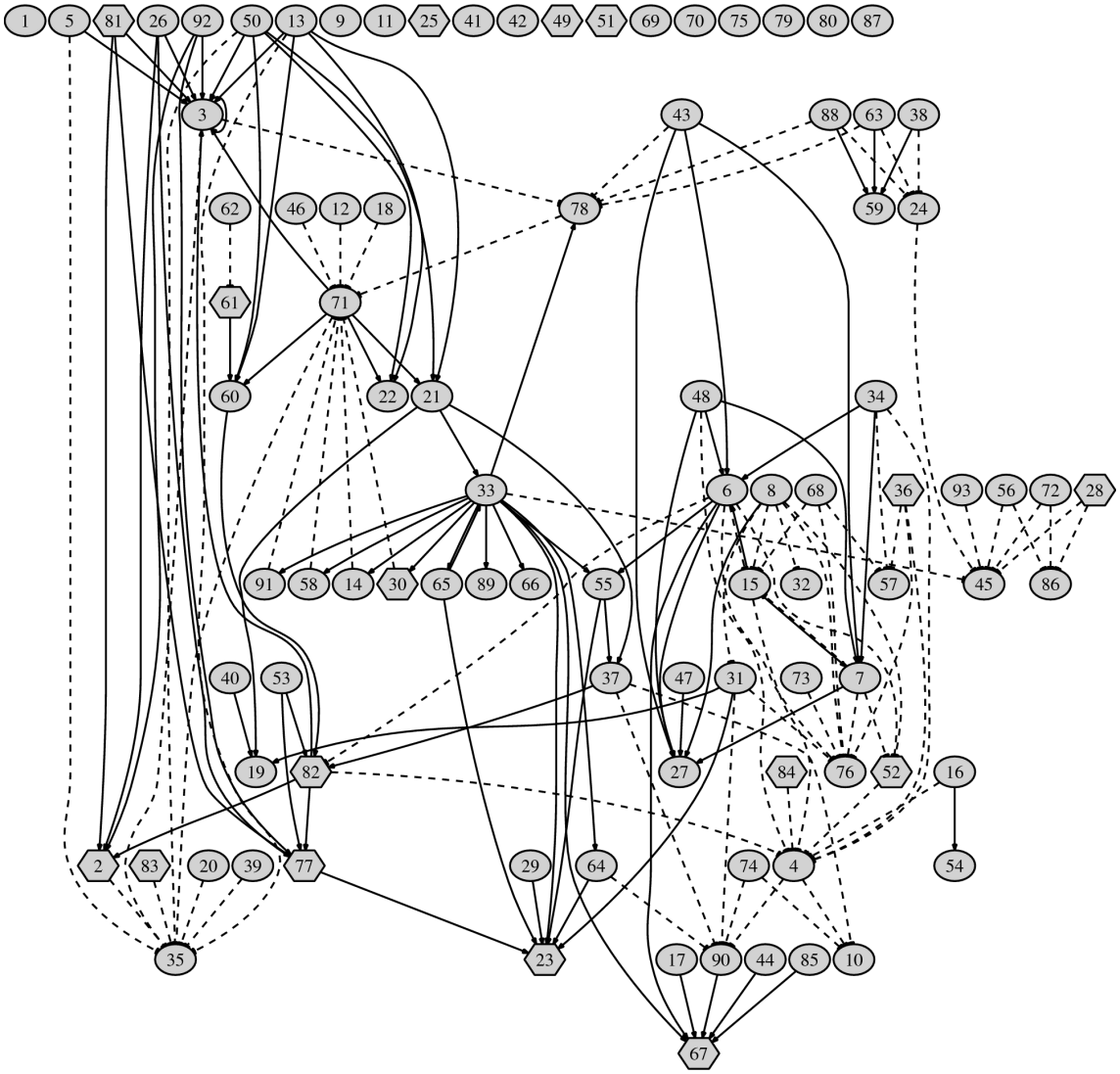


Figure III.5 – (from Beslon et al., 2010b). Regulation network of the final best organism of a representative simulation with low mutation/rearrangement rates. After 15,000 generations of evolution with $\mu_{mr} = 5 \times 10^{-6}$, the best organism in the population presents a relatively large and complex regulation network (93 genes and 73 TFs, 13 of which being pure TFs). Solid lines represent activation links and dashed lines, inhibition links. Genes having a metabolic activity are represented by ellipses. Hexagons represent genes without any metabolic activity. Genes that have a regulatory activity (outgoing edge) are transcription factors (TFS). Then hexagons with outgoing edges are pure TFS. This network was generated using the graphviz software.

The question of the origin and universality of such scaling laws remains open (Cordero and Hogeweg, 2007; Molina and van Nimwegen, 2009). Some evolutionary models based on gene duplication and deletion can produce power-law relations (Luscombe et al., 2002; Foster et al., 2006) but these models directly consider the mutations that went to fixation

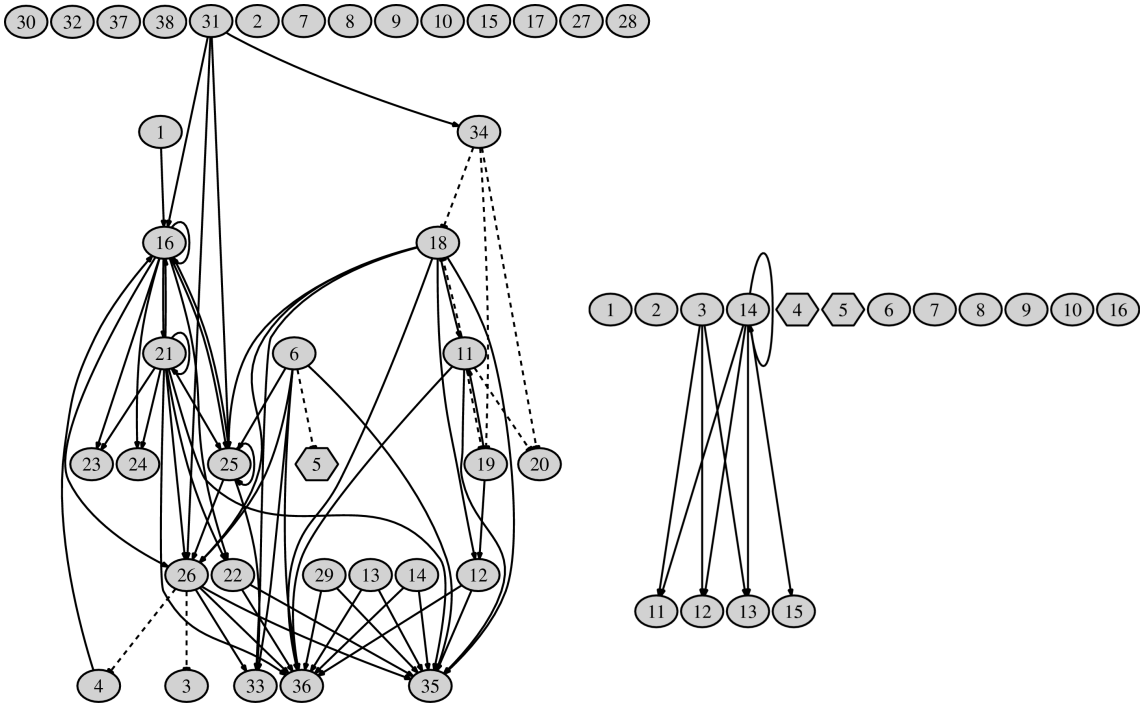
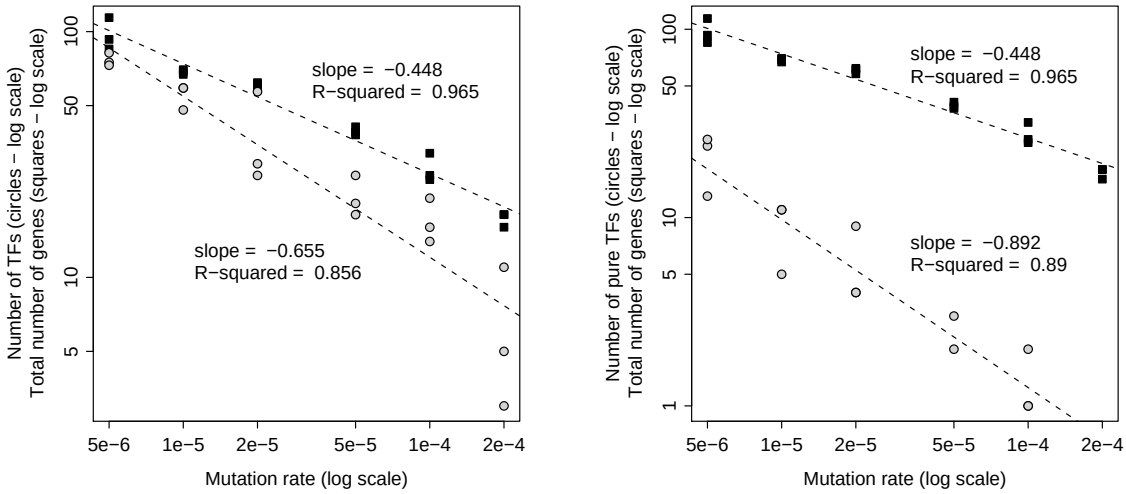


Figure III.6 – (from Beslon et al., 2010b). Regulation networks of the final best organisms of a representative simulation with respectively moderate **(left)** and high **(right)** mutation/rearrangement rates. After 15,000 generations of evolution with respectively $\mu_{mr} = 5 \times 10^{-5}$ and $\mu_{mr} = 2 \times 10^{-4}$, these organisms present networks of respectively medium complexity (left – 38 genes and 18 TFs) and low complexity (right – 16 genes and 2 TFs). Solid lines represent activation links and dashed lines, inhibition links. Genes having a metabolic activity are represented by ellipses. Hexagons represent genes without any metabolic activity. Genes that have a regulatory activity (outgoing edge) are transcription factors (TFS). Then hexagons with outgoing edges are pure TFS. These networks were generated using the graphviz software.

in the population, without distinguishing the respective influences of the various underlying processes (genetic drift, natural selection, mutational biases). However, the classical hypothesis is that the scaling has a selective origin. It is often assumed that these scaling laws result from a selection process depending on the organisms' lifestyle: complex environments would require the coordination of multiple metabolic pathways (Cases et al., 2003). Alternatively, it was argued that any increase in the genetic repertoire of an organism (e.g., a new metabolic pathway) generates a need for new transcription factors in order to regulate its activity within the existing metabolism (Maslov et al., 2009).
Actually, despite the tremendous advance in the fields of genomics and transcriptomics, it is still not clear whether these scaling laws result from selective constraints (e.g., selection for integrated networks), from the intrinsic dynamics of the evolutionary process or from any other mechanism still to be revealed (Molina and van Nimwegen, 2009).
As shown in figure III.9, our experiments with R-Aevol reproduce qualitatively the scaling laws observed in the prokaryotic kingdom (Cases et al., 2003; van Nimwegen, 2003; Konstantinidis and Tiedje, 2004; Molina and van Nimwegen, 2008). At one extreme, small



(a) All TFs (b) Pure TFs

Figure III.7 – (from Beslon et al., 2010b). Number of genes having a metabolic activity for the final best organisms of each simulation (black squares). Grey circles represent either the number of transcription factors **(a)** or the number of pure transcription factors **(b)**. Both values clearly scale with the mutation rate but the number of TFs grows faster than the number of genes.

genomes with only a few genes hardly have any regulation connections between the genes they bear. At the other extreme however, large genomes result in large and complex regulation networks within which many genes interact with one another in a very intricate way. Both the number of metabolic genes and of TF-coding genes scale as power-laws with the total number of genes. However, while the former scales at most linearly, the latter shows a super-linear scaling – figure III.9(a) –, going up to a quadratic scaling when considering pure TFs – figure III.9(b).

Since, in our experiments, all the organisms evolved in an identical steady environment, the differences we observed in the complexity of either the genomes or the regulation networks cannot have been caused by environmental conditions. The only differences throughout our simulations was the mutation rate, that ranged from a very high rate ($\mu_{mr} = 2 \times 10^{-4}$) to a low one ($\mu_{mr} = 5 \times 10^{-6}$). As figures III.3 through III.9 show, the mutation rate is a strong determinant of the complexity of an organism. Using the Aevol model, it has already been shown that this scaling of the genome size, amount of non-coding DNA and number of genes with respect to the mutation/rearrangement rate are the consequence of an indirect selection of those lineages whose genomic structure allow for an appropriate level of mutational robustness (see (Knibbe et al., 2007a), chapter I, section 9.2 and chapter II). Figure III.10 shows the fraction F_ν of neutral offspring of the final best individual of each R-Aevol simulation. As in the previous experiments, this fraction seems to be close to that leading to the production of one neutral offspring per

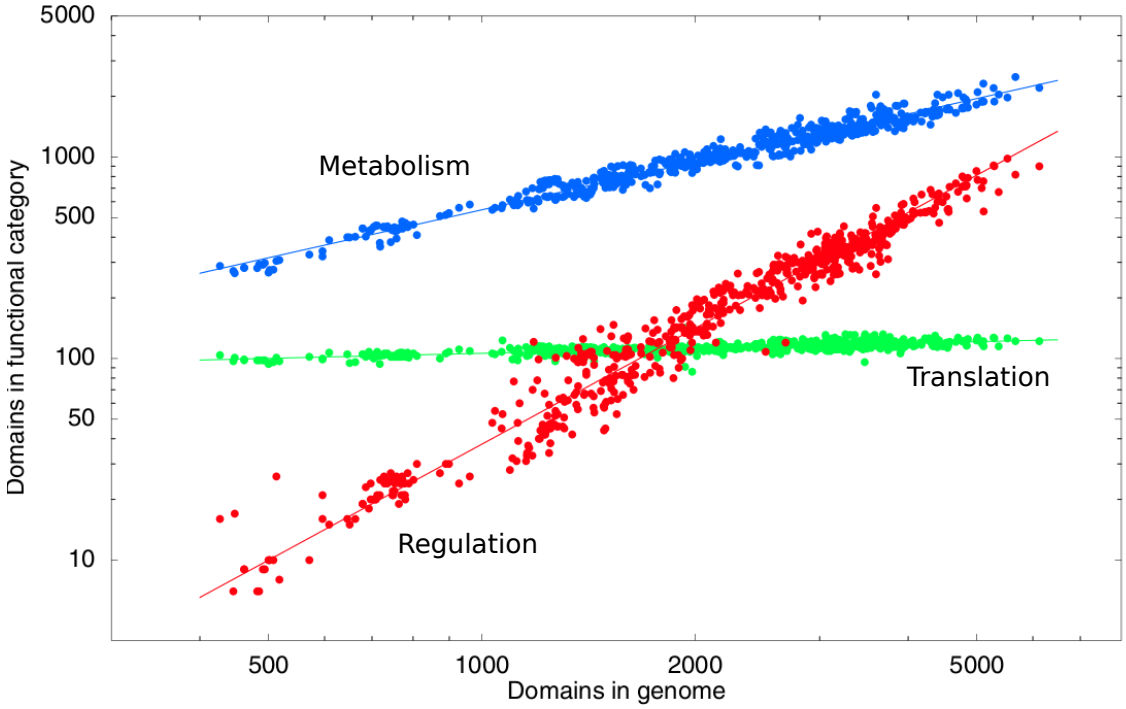
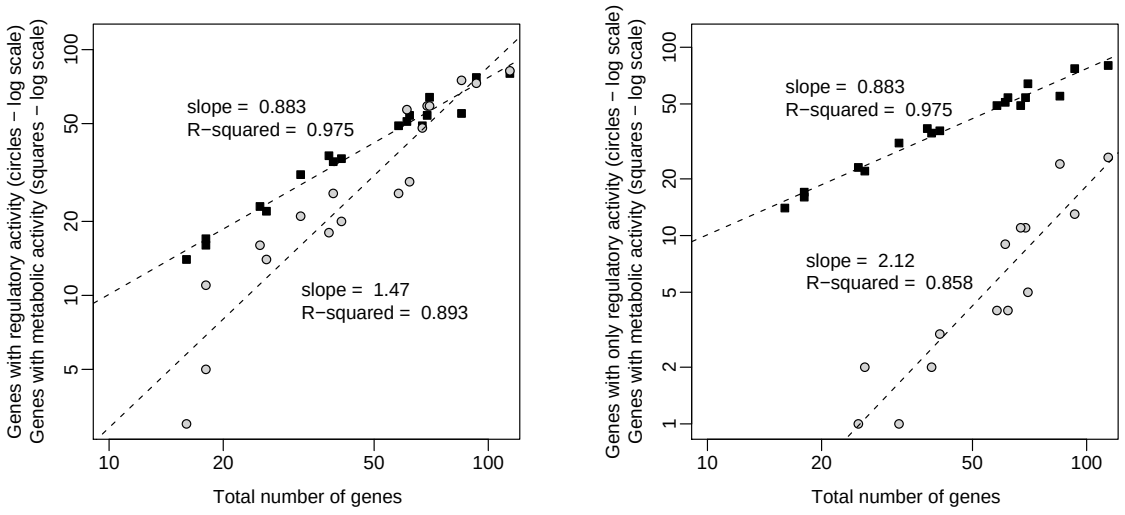


Figure III.8 – (from Molina and van Nimwegen, 2008). Number of protein-domains associated with functional categories translation, metabolism and regulation as a function of the total number of domains in the genome for which a functional annotation is available. Each dot corresponds to a fully-sequenced microbial genome.

generation ($F_\nu W \approx 1$ with W the average number of offspring of the best individual in the population), although for very low values of μ_{mr}, F_ν seems to be under this value of $\frac{1}{W}$. These low values of F_ν could be explained by a high homogeneity of the population, but further tools will need to be developed for this hypothesis to be asserted.

In fact, all the scaling laws observed in this experiment stem from this long-term pressure, and more specifically on that exerted on the number of genes. Indeed, as the number of genes is drawn up, the number of promoters also increases [1], resulting in a super-linear growth of the number of putative gene-promoter associations. Because in the model, the regulatory activity of a protein is computed through a combinatorial algorithm that associates protein primary sequences with the sequences of promoter binding sites, any increase in the number of coding RNAs (RNAs containing at least one gene) comes along with an increase in the number of potential targets for each protein bearing a regulation domain. Indeed, each TF binding site being a 20-bp long sequence, the probability of having exactly the same sequence for different binding site is quite low, even when several promoters are homologous, created by duplication-divergence, because they diverge very quickly. The consequence of this increase in potential targets is that any protein containing at least one regulation domain has a greater probability of actually regulate something in

1. Even though it was not measured in this specific experiment, the results presented in chapter II suggest that the number of RNAs bearing at least one gene should increase slightly faster than the number of genes when the mutation rate is lowered.



(a) All TFs

(b) Pure TFs

Figure III.9 – (from Beslon et al., 2010b). Number of genes involved in metabolism (black squares) and in the regulation process (grey circles) as a function of the total number of genes in the genome (best organisms of the 18 simulations at generation 15,000). Grey circles represent either the number of transcription factors **(a)** or the number of pure transcription factors **(b)**. Dash lines show power-law fits.

a large genome than in a small one.

R-Aevol thence appears as a null model in which links in the networks are added with an almost constant probability when the number of gene-promoter pairs increases[1]. Consequently, in this model, the scaling of the number of genes because of mutational robustness constraints naturally leads to a super-linear increase in the number of regulatory nodes.

Whether a similar mechanism can explain the quadratic growth of Transcription Factors observed by van Nimwegen (2003) and Molina and van Nimwegen (2008) is an open question. Since real transcription factors have one or more DNA-binding domains that are well defined units on the structural, functional and evolutionary level, it is not clear whether such combinatorial process is at work in real genomes. Yet, several authors have reported the combinatorial properties of the binding between TFs and their DNA targets. According to Itzkovitz et al. (2006), the number of degrees of freedom of the binding mechanism can partly account for the increase in the number of TFs. Moreover, it is also known that TFs can bind to a broad spectrum of binding sites with different affinities and change targets widely among species (Balleza et al., 2009).

Maybe the most striking result of our simulations is that the super-linear growth of the

1. At one stage, if a great variety of the possible binding site sequences are present in the genome, this process should reach a saturation point. Then adding a promoter would be less likely to cause the creation of a new target sequence. However, this "saturation point" is much higher than the number of genes/promoters we observed in our simulations.

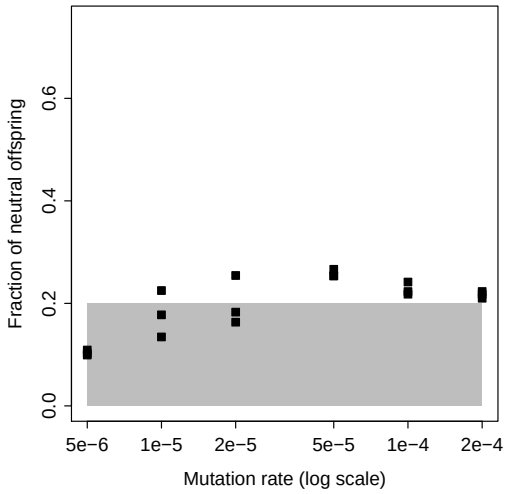


Figure III.10 – (from Beslon et al., 2010b). Fraction F_ν of neutral offspring of the final best organism for each simulation. These values were estimated by forcing the final best organism of each run to reproduce itself 10,000 times (with the same mutation rate as during the run) and by counting the number of offspring that have the same fitness as their progenitor. The grey area represents organisms whose fraction of neutral offspring is lower than the value leading to an average of one neutral offspring per generation.

number of TFs is also observed for *pure* TFs. Indeed, pure TFs do not directly contribute to the metabolism, their only impact on the phenotype of the organism happening through the variations they can cause in the concentration of other proteins. Since, in these experiments, the environment is constant over time, an organism could be well adapted with no regulation at all. It is hence surprising to observe a significant number of pure TFs in these simulations and even more that their number scales super-linearly with the number of genes. In fact, pure TFs scale more than quadratically with the number of genes (figure III.9). One can propose different hypotheses to explain the appearance and fixation of pure TFs. They can appear due to random mutations but they most likely result from duplication/divergence events (e.g., gene copies that lose their metabolic activity while retaining their regulation activity). The interesting question is why evolution maintains such genes in the simple environment where our organisms live. One can assume that, when the number of genes increases, there is a need for more regulation in order to position the attractor of the network more precisely in a space in which the number of dimensions increases (Maslov et al., 2009). In this hypothesis, pure TFs could be directly selected for. Alternatively, one can suppose that they are indirectly selected; however, their contribution to the robustness/evolvability balance is very difficult to assess. They can contribute to the organism's robustness if they have a canalizing effect. They can also contribute to the organism's evolvability by enabling small mutational variations that may be more likely to be positive than mutations in metabolic genes. In this hypothesis,

pure TFs would be conserved because their mutation can finely tune the activity of their target proteins without changing the metabolic processes these targets are involved in.

4 Gene Regulation Networks in a Complex Environment

The results we have presented in the previous section show that, in a simple case, *i.e.* in a steady environment in which regulation is not mandatory and hence the regulation network itself is incidental, the mutation rate is a strong determinant of the size and complexity of both the genome and the regulation network. The extension of this result to complex environments is still to be done. Indeed, a full campaign of experiments using the R-Aevol model in complex environments would require the R-Aevol code to be optimized, which is currently under progress. Yet, we conducted a limited experiment using a single set of parameters to explore the structuration of gene regulation networks in the case of non-trivial environments. The results show that the evolved networks are much more complex than is necessary. In fact, the complexity of the network was such that we had to use data-mining tools to unravel its dynamics. This suggests that, as is the case in trivial environments, the complexity of the network is at least partly driven by indirect pressure towards a specific level of mutational variability of the phenotype.
Here, we present this simple experiment and describe the procedures we used to decipher the dynamics of the complex regulation network that evolved in the model. We conducted gene knock-outs experiments, a method inspired by "wet" biology, to produce data that could subsequently be analyzed by a data-mining algorithm in order to help us understand the behaviour of the network.

4.1 Experimental Setup

Using the R-Aevol model, we let 3 different populations of 1,000 individuals evolve for 40,000 generations under a moderate mutation/rearrangement rate ($\mu_{mr} = 10^{-5}$ per mutation type per base-pair) and a mild selective pressure in a dynamic environment. This environment (identical throughout generations) was divided into two periods of 10 time steps each within the lifetime of an organism. During the first period of 10 time-steps, the environment was identical to the one we used in all the previous experiments presented in this thesis (see figure I.2), then, during the second period, the environmental target was modified, one of the lobes of the environment being removed, and an external signalling protein was introduced in the system so that the virtual organisms can "sense" the environmental change. Organisms were evaluated twice (once at the end of each period), meaning that, in order to be well-adapted, the organisms *must* evolve a reaction to the environmental perturbation through an effective regulation network connected to the signalling protein. Figure III.11 shows a typical phenotype evolved under such conditions. The values we used in these experiments for the main parameters are presented in table III.2.
In all three simulations (which will henceforth be referred to as S1, S2 and S3), the organisms progressively acquire new genes that are connected in the regulation network in such a way that the proteome fulfils the regulation task the organisms are selected for

Parameter	Value
N	1,000
nb_gener	40,000
$init\_length$	5,000
$init\_method$	Clonal, One Good Gene
$selection\_scheme$	Linear Ranking
η^+	1.998
$E = \sum_i \alpha_i G_i$	$\alpha_1 = 1.2; G_1 : \mu = 0.52; \sigma^2 = 0.12$
	$\alpha_2 = -1.4; G_2 : \mu = 0.2; \sigma^2 = 0.07$
	$\alpha_3 = 0.3; G_3 : \mu = 0.8; \sigma^2 = 0.03$
$env\_sampling$	300
μ_{point}	$\mu_{mr} = 10^{-5}$
μ_{s_ins}	
μ_{s_del}	
μ_{dupl}	
μ_{del}	
μ_{inv}	
μ_{trans}	
max_indel_size	6
W_{max}	0.033333333

Table III.2 – Parameters used in the experiments presented in this section. Mutation and rearrangement rates take their values among those proposed, one common value for each types of operators.

(figure III.12). After 40,000 generations, we focus on the best individual of each run. They all show complex regulation networks with 51 (S1), 34 (S2) and 58 (S3) genes respectively. These genes are connected by hundreds of links, but while the number of genes is only slightly variable from one simulation to the other, the number of connections strongly differs: network S1 has 328 connections, S2, 153 and S3, 908. Note that there is no direct correlation between the size of the network and the metabolic error: S2, having the smallest network, has a better fitness than S1 which is much larger (figure III.12). Figure III.14 (squared part) shows the variation of protein concentration of the best final individual in S1 [1]. One can easily see that, at time $t = 10$ (i.e., when the signalling protein is introduced) and $t = 20$ (i.e., when this signal is switched off), the protein concentrations quickly change to stabilize on new values. When looking at the phenotype (figure III.11), we see that the signalling protein triggers a reorganization of the phenotype, one lobe of the phenotype vanishing from $t = 10$ to $t = 20$, thus following the environmental change. When looking at the genetic networks created by evolution (figure III.13), the complexity of the network is striking and is way over that needed to complete the task (the environment is made up of only two states and the organisms need not be bistable since the external protein is present throughout the whole second period). Again, the complexity

1. These measures were obtained by simulating the life of the organism during 30 time steps, the external signal being introduced at $t = 10$ (as it was during the evolution), and taken away at $t = 20$.

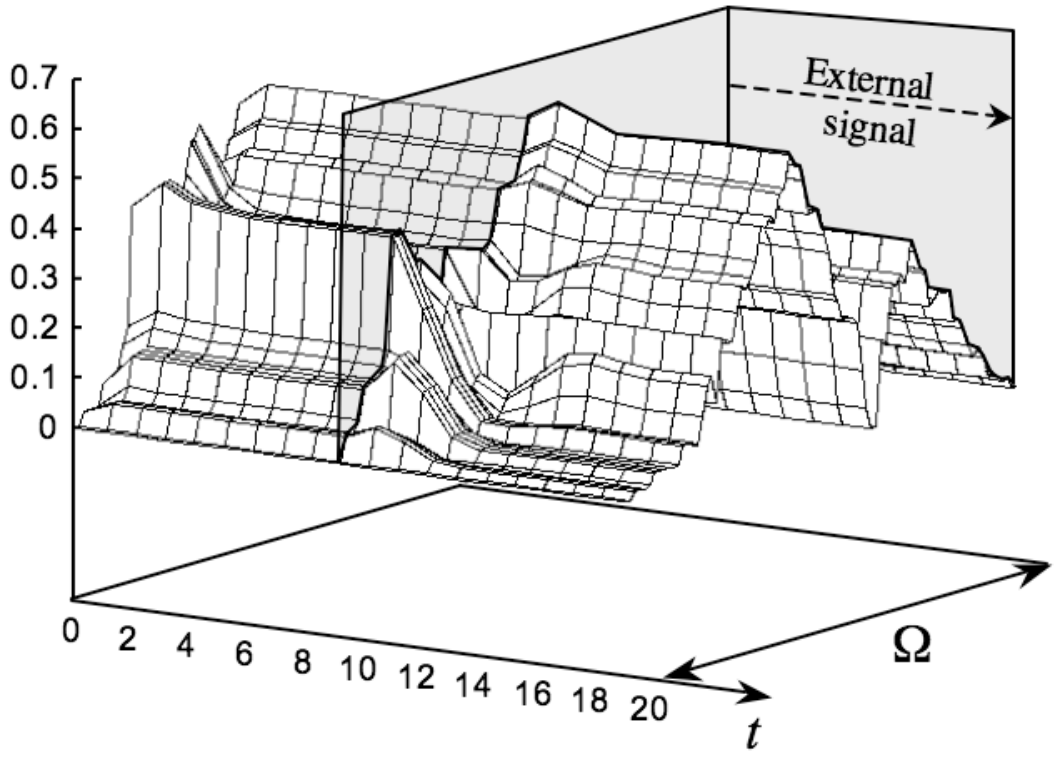


Figure III.11 – (from Beslon et al., 2010a). At each time step t the phenotype is expressed as the efficacy of the organism in performing the metabolic functions in the abstract set Ω. Here, at time $t = 10$, an external signal is sent to the organism which reacts by modifying its metabolic profile. The metabolic error is measured at $t = 10$ and $t = 20$

of the network appears to be mainly driven by the rates of mutation and rearrangement. Now, it clearly appears that the dynamic behaviour of the network is impossible to decipher directly from the network structure. Thus, when analyzing the results of our experiments, we face the same problem as practitioners do with real biological networks: we need automatic mining algorithms to help us understand the structure of the networks and link it to their dynamic behaviours.
Going one step beyond, we decided to perform, with our *in silico* organisms, the same experiments biologists do with real bacteria: we generated mutant variants in which single genes are invalidated one at a time (KO-mutants, section 4.2) and measured their transcriptional activity. The resulting dynamic data can then be analyzed to understand the structure of the regulation network using data-mining algorithms. Following this idea, we propose to use this data as a benchmark, available for the scientific community to test knowledge discovery algorithms [1].

4.2 Gene Knock-Outs in Digital Models

Gene Knock-Out (KO) (Galli-Taliadoros et al., 1995) is a widely used technique in molecular biology. It provides geneticists with an insight into complex mechanisms, focusing

1. All data are available on Internet: http://liris.cnrs.fr/guillaume.beslon/IDAj.data/

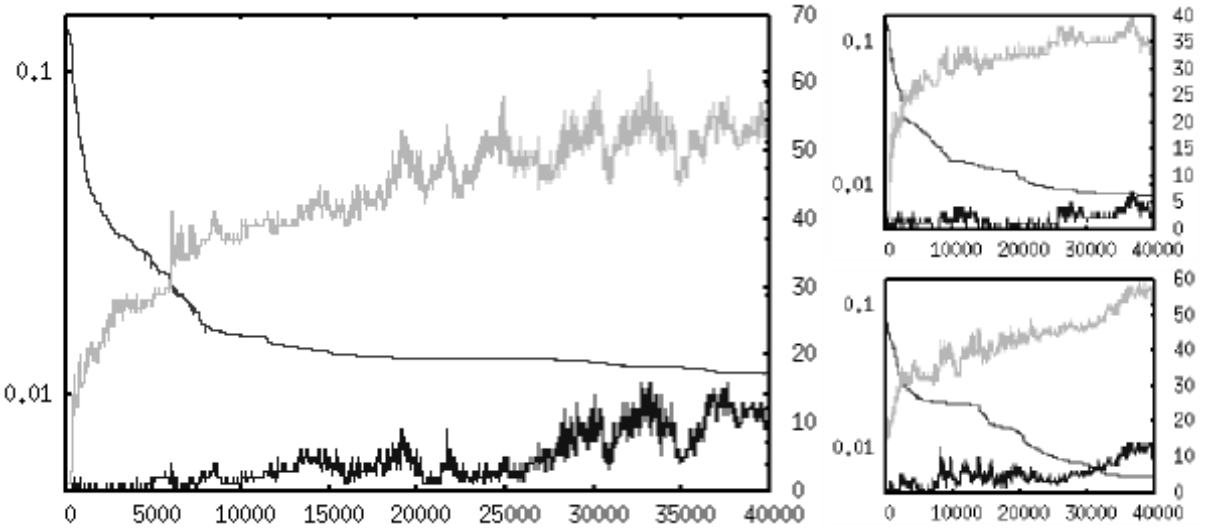


Figure III.12 – (from Beslon et al., 2010a). Evolution of the organisms during 40000 generations. Left: Simulation S1. Right: Simulations S2 (top) and S3 (bottom). Left axis: metabolic error (black decreasing line). Right axis: number of genes (light grey) and genes which do not contribute to the metabolic activity (dark grey). The increasing number of non-metabolic genes in the second stage of the evolutionary process is characteristic of the recruitment of pure Transcription Factors (TFs).

on the contribution of a particular gene or set of genes. It consists in producing a mutant lineage in which a targeted gene is invalidated ("knocked-out"), thus preventing the corresponding protein from being produced by the organism under study. The phenotype of the mutant organism is then compared to the original one (the "wild-type") with the objective of understanding the role of the knocked-out gene in the organism.
Using KO mutants, both the direct phenotypic contribution of a given protein and its role in the regulation network can be studied. To study the network, one needs to focus on transcriptome data that gives information about the expression of genes. Systematically knocking out every gene within a genome allows the geneticists to carry out broad comparative studies that could shed light on the complex structure of gene networks (Gu et al., 2003). Indeed, knock-outs can be used to create perturbations on a gene network in order to help infer its hidden structure (Hecker et al., 2009). However, the systematic knock-out of genes in any genome, even the smallest, yields a vast amount of data that is very difficult to process "by hand". It is hence necessary to develop data mining tools that can help analyse knock-out data (Geier et al., 2007).
In silico models have the advantage of providing us with access to any piece of data we might need in order to understand the system's behaviour, even after the experiment is finished. They are hence particularly suited to generate benchmark datasets for data mining methods. We simulated a systematic knock-out process on the evolved organism from simulation S1, generating time series representing the concentration of proteins over time for each gene's knock-out. We measured these concentrations during 30 time-steps, the external signal being present during the middle 10 time steps. Figures III.14 and III.15 show some examples of mutant behaviours represented by variations of protein

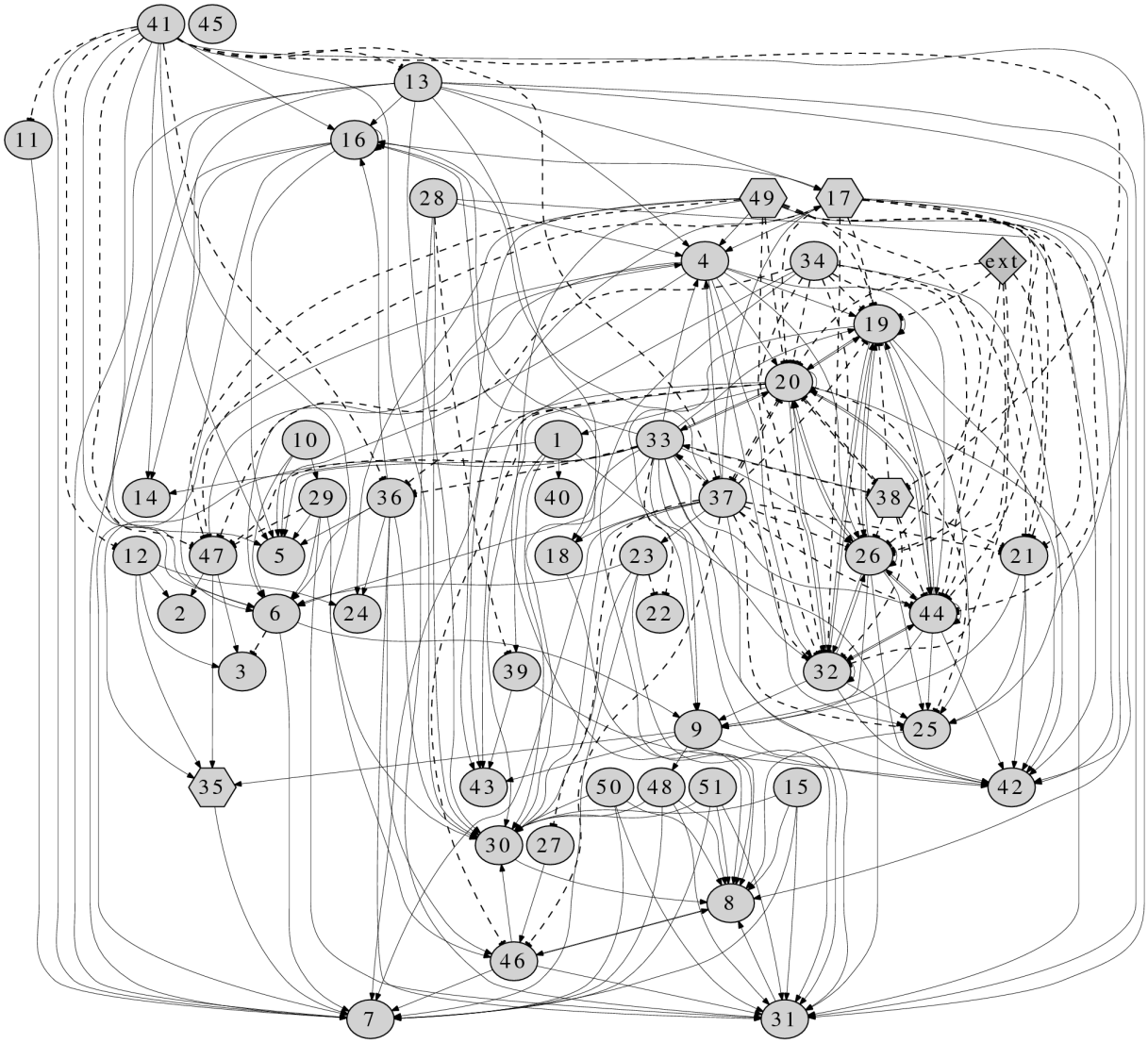


Figure III.13 – (from Beslon et al., 2010a). Structure of the regulatory network after 40,000 generations (experiment S1). The network contains 51 genes, 4 of which being pure transcription factors (hexagons, genes 17, 35, 38, 49), the remaining all having a metabolic activity (ellipses). The signalling protein (dark grey diamond) is connected to the network through a complex connectivity pattern whose behaviour is quite impossible to decipher manually. Solid lines represent activation links. Dashed lines represent inhibition links. This network was generated using the graphviz software.

concentrations (figure III.14) and of the phenotype (figure III.15). This set of data was then used to understand the underlying gene regulation network thanks to an *ad-hoc* mining algorithm.

4.3 Mining the KO sequences

A simple data-mining algorithm was developed by Christophe Rigotti as a test case for the usability of our gene KO data for benchmarking purposes. As it is outside the scope of

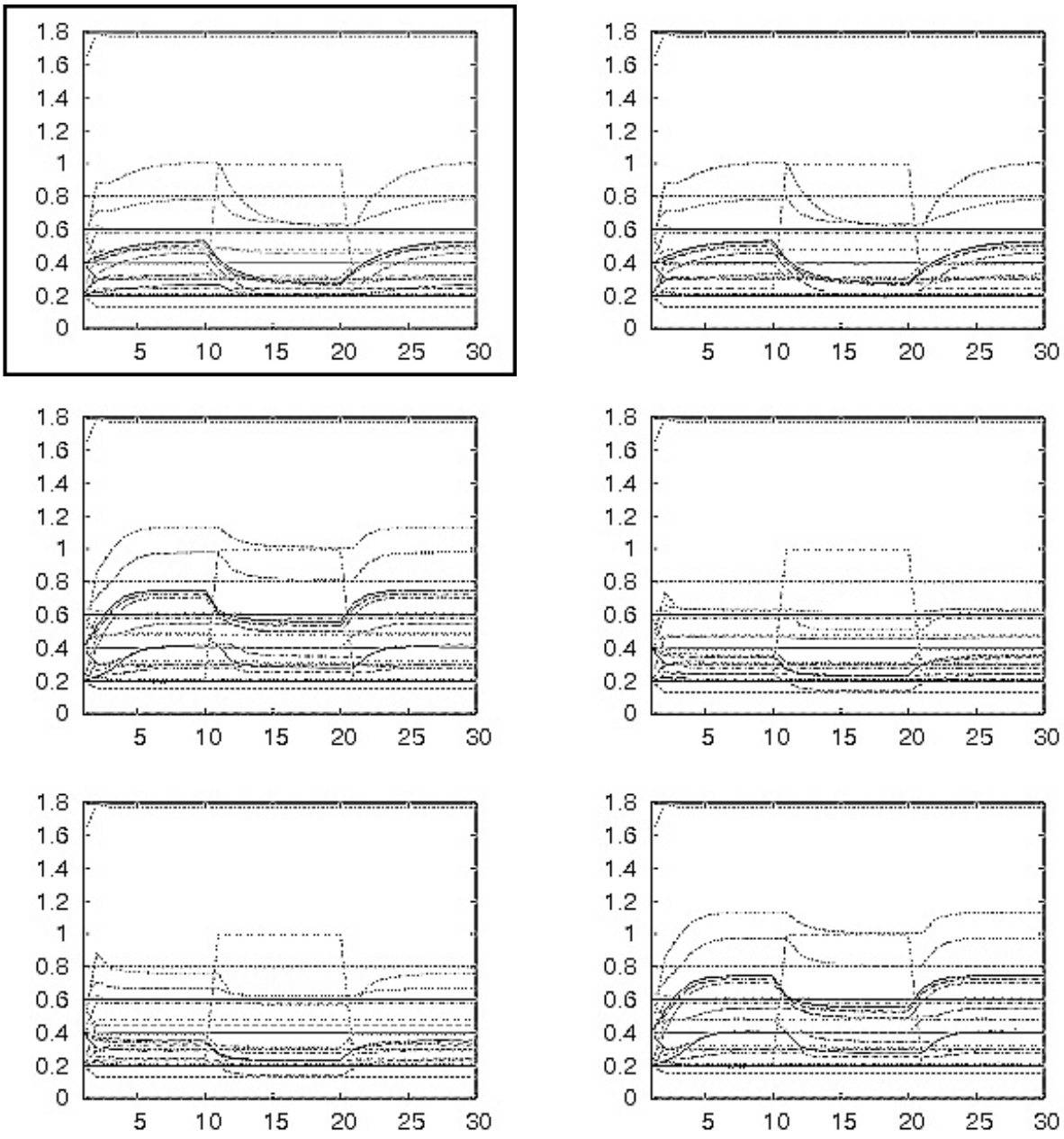


Figure III.14 – (from Beslon et al., 2010a). Variations of the concentration of each protein over time in the wild-type organism of S1 (squared) and in five KO mutants. From top to bottom and left to right: wild type, KO of gene 1, KO of gene 17, KO of gene 19, KO of gene 20 and KO of gene 34. x axis: organism's life time (time steps). y axis: protein concentrations (arbitrary units). Displayed mutants have been chosen because they show clear differences when compared to the wild-type.

this thesis, I will provide here only a brief description of this algorithm; for more details, please refer to Beslon et al. (2010a).

The objective of this method is to provide the expert with some information about genes having similar KO effects. More precisely, the method aims at exhibiting groups of genes

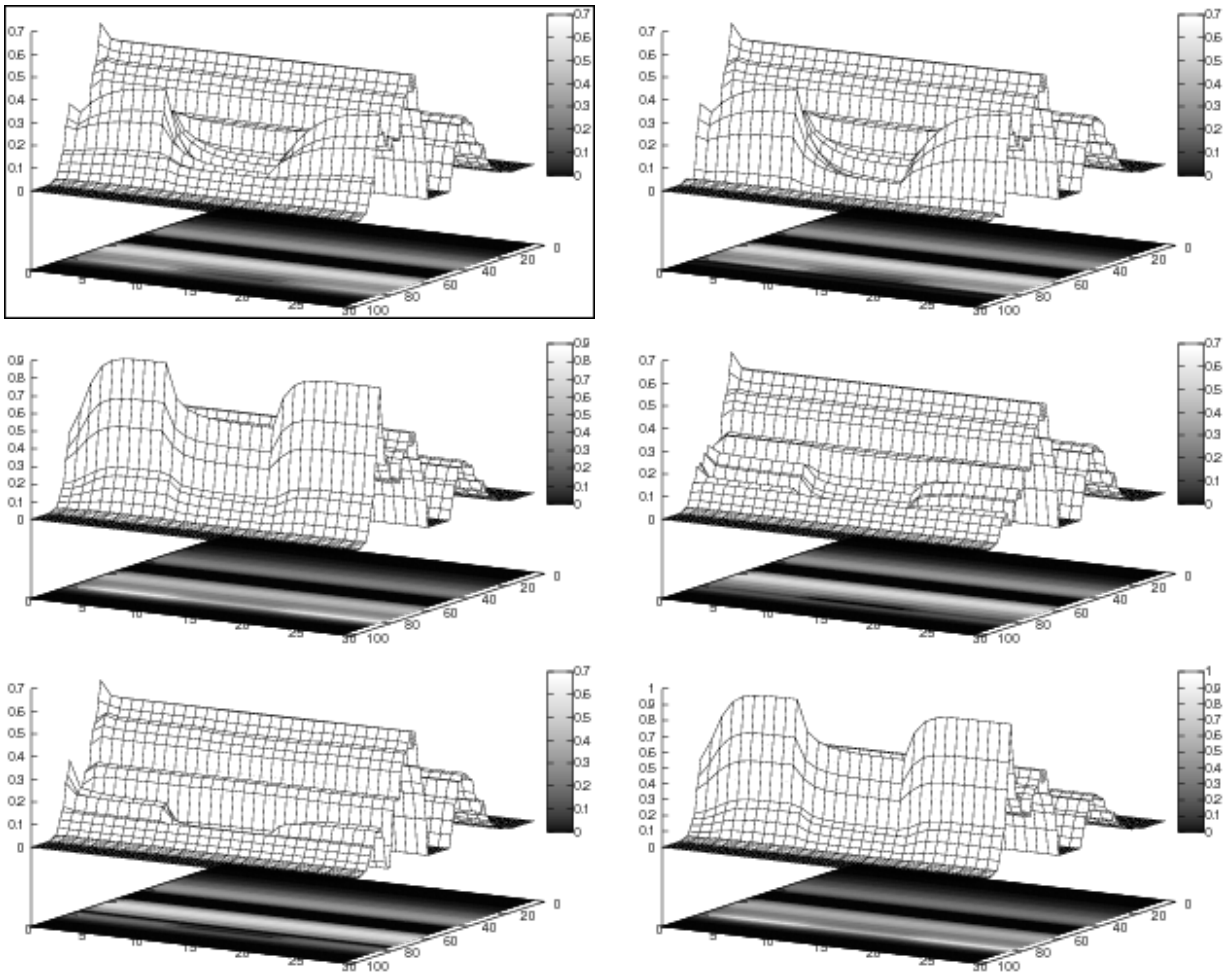


Figure III.15 – (from Beslon et al., 2010a). Phenotypes of the wild-type organism of S1 (squared) and of five KO mutants. x axis: life time of the organisms (time steps). y axis: functional space. z axis: efficacy of the organism in performing the function. Mutants are the same as those shown on figure III.14 (KO of genes 1, 17, 19, 20 and 34).

whose KO lead to similar changes in protein concentration values. This information is obviously useful if we do not know the regulation network, to suggest genes that are involved in the same regulation process. But such groups are interesting even when one has access to the regulation network as it is the case here. Indeed, the intricacy of the evolved networks makes it impossible to consider them frontally. In this case, identifying groups of genes having a similar effect on the transcriptional patterns can help the expert to gain some understanding of the underlying regulation processes.

The data describes a set of experiments, each of which corresponds to the KO of one single gene. For each KO, we recorded the concentration of all the proteins during the whole life of the organism ($m = 30$ time steps). We also have at hand an additional experiment used as a reference, where no KO is performed. This experiment using the "wild type" organism provides us with the "normal" protein concentrations. Let the genes be numbered from 1 to n and let us denote exp_i, with $i \in [1, n]$, the data obtained for the KO of gene i. Then

exp_i can be represented as a $n \times m$ matrix of concentration values $c_{j,t}$, where $c_{j,t}$ denotes the concentration of protein j at time t. Similarly, the wild type experiment denoted exp_{wt} is also represented as a $n \times m$ matrix of concentration values.
The global mining process contains 3 main steps: discretization, identification of values of interest (within the concentration values) and extraction of groups of genes having similar patterns regarding the identified values of interest.

Discretizing and identifying values of interest First of all, the protein concentration values of each organism (wild type and KO mutants) are discretized into concentration levels, the boundaries of which are determined after a quick visual inspection of the data. Each element $c_{j,t}$ of each matrix exp_i (and of exp_{wt}) is then replaced by a discrete label of the form $(j, f_d(c_{j,t}))$ where f_d is the mapping function we use for the data discretization. Then, given a threshold α, we retain as *main concentration values* (values of interest), only the labels $(j, f_d(c_{j,t}))$ that appear both in the wild type experiment exp_{wt} and in at least α percent of the KO experiments. The set of main concentration values thus obtained is denoted $\mathcal{M}$.

Finding groups of genes Let p be the number of main concentration values (i.e., the size of $\mathcal{M}$). We define the $p \times n$ Boolean matrix L, each element $l_{k,i}$ of which is set to one if the k^{th} main concentration value has disappeared in the KO of gene i (recall that every main concentration value is by definition present in the wild type), and is set to zero otherwise. Note that no information is retained regarding the time of occurrence of these main concentration values. Then, given an integer threshold σ, we extract the sets of genes such that the KO of these genes have in common at least σ missing main concentration values.
Finally, since we are interested in sets of genes that contain many genes and that share many missing main concentration values, we assigned a rank to each of these groups depending on both of these criteria.

4.4 Results

This mining process has been applied to the KO data obtained from the final best organism of simulation S1, whose network is shown in figure III.13. Our data-mining process identified 75 groups of genes, most of which were actually subgroups of other – better ranked – groups. After elimination of these subgroups, the focus was clearly put on two groups of genes: $\{4, 19, 20, 26, 32, 38, 44\}$ (group 1) and $\{17, 34, 37, 38, 49\}$ (group 2). Figure III.16 shows a zoom on these groups taken from the original network (figure III.13).
Using these two groups as entry-points into the otherwise very difficult to understand network, we were able to fairly easily understand the way this network works and how it manages to respond to the appearance of the external signal. Indeed, looking closely at the outgoing edges of the genes from groups 1 and 2, we can see that a subgroup of group 1 (genes 19, 20, 26, 32 and 44) is actually a clique, each gene from this subgroup activating the others, Genes in group 2 however are not connected to one another. Yet, they do share a common behaviour: they all inhibit the genes of the clique we have identified in group 1 and most of them activate gene 4, which in turn activates genes from the identified clique. These observations enabled us to draw a sketch of the network in a way that is meaningful

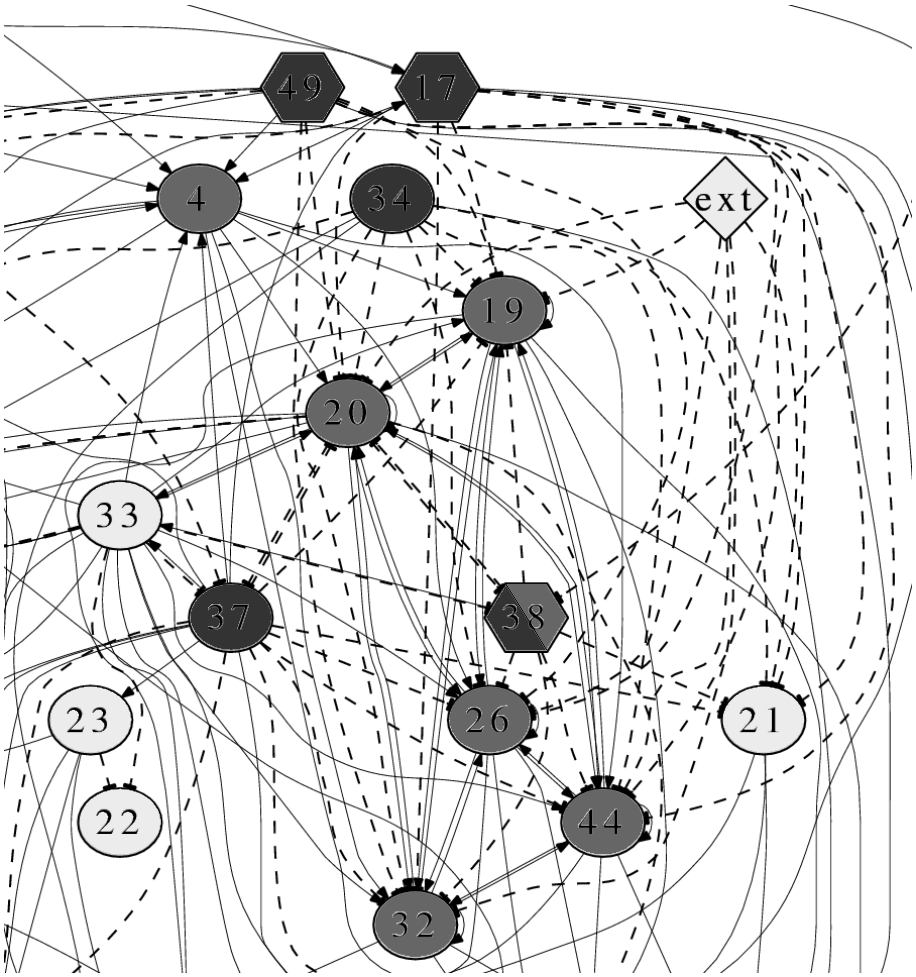


Figure III.16 – Zoom on the groups of genes that were identified by the data-mining algorithm. Genes from group 1 are shown in dark grey, genes from group 2 in darker grey. Genes showed in light grey correspond to neither group. Note that gene 38 belongs to both groups.

to a human being (figure III.17). It is basically made up of two functional modules that we were able to derive from the groups given by our mining process. The first module (module A: genes 19, 20, 26, 32 and 44) is the clique we have identified within group 1. The proteins corresponding to these genes all code for a metabolic activity in the lobe that has to be turned off when the external signal appears. This module displays a motif that is similar to a positive-feedback-loop (Alon, 2007). This module is directly inhibited by the external protein as well as by the second module (module B: genes 17, 34, 37 and 49) which is a subset of group 2. Module B has a double negative effect on module A: not only does it down-regulate A but the metabolic contribution of the proteins from B is the repression of the very metabolic functions activated by proteins from A. The positive regulation of module B on module A through gene number 4 seems incoherent but it probably helps the organism to fine-tune the expression of the genes of module A[1]. It

1. This gene and the two modules form an incoherent Feed-Forward Loop – iFFL. Such a loop is known to accelerate state switch (Alon, 2007). However, since in our case the organism is not evaluated

is important to note that, while we might have been able to detect module A "by hand" because it is a clique, we would surely not have discovered module B since its constituting nodes are not connected with one another. We are facing a functional module whose unity is important for the organism but which does not correspond to any structural module in the network. Such a module would obviously have been missed by any link-pruning module detection algorithm (Newman and Girvan, 2004). The emergence of such modules during the evolution process, as well as its discovery by our KO-mining algorithm was a surprise. We now plan to conduct similar experiments to test whether or not this type of structure emerges repeatedly.

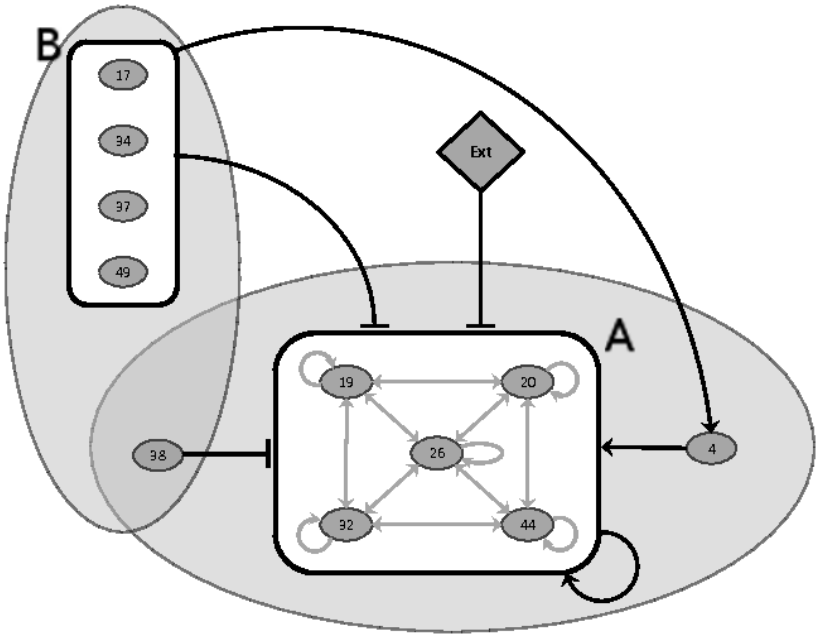


Figure III.17 – (from Beslon et al., 2010a). Sketch view of the network of experiment S1 (see figure III.13 for the full network). The two groups identified by the mining algorithm are depicted by the grey ellipses. Module A (bottom) is composed of five genes forming a positive-feedback-loop motif. Module B (left) is composed of four genes that are not connected with one another: These genes fulfil the same role in the network although they are not interconnected. Gene 38 can be considered as belonging to the same module even though it does not activate gene 4. The external signal (grey diamond) triggers the activity of module A. Vee-Arrows: Activations; Tee-Arrows: Inhibitions.

All the other genes in this network have little impact on the dynamics of the network. Of course, knocking-out any metabolic gene will cause the metabolic functions it was involved in to be either under- or over-realized depending on whether the gene was an activating or an inhibiting gene. But the impact of these KOs on the concentration levels of other proteins was limited, suggesting that their connection to the network was mainly incidental.

until ten time steps has elapsed, meaning the network has ample time to switch, it is not clear whether the iFFL has been selected to decrease the switch delay when the signal is perceived or simply to precisely tune the module activity.

5 Conclusion

In this chapter, we have presented an extension of the Aevol model, R-Aevol, in which an explicit process of regulation of gene expression was introduced (Sanchez-Dehesa, 2009). We then presented two sets of experiments we conducted with this model. The first set of experiments we presented was conducted in a simple steady environment in which regulation was not mandatory for an organism to be well adapted.

Our results show that the model still reproduces scaling laws observed in real organisms at the genomic level. Furthermore, we observed that evolution produced organisms with regulation networks way more complex than what is required to fulfil the task they were selected for. Interestingly, the complexity of the regulation network itself seems to be governed by the same parameters that shape the genome structure: the mutation rate and the rearrangement rate. We showed that these scaling laws reflect fundamental principles of bacterial evolution, namely the selection for an appropriate balance between robustness and evolvability (Lenski et al., 2006). Indeed, our simulations show that the pressure for complexification of the network can be indirect, unrelated to differences in the environment or the organisms' lifestyle: even when facing identical environmental constraints, the evolved structure can range from very simple life forms (with a reduced gene set and loose connectivity) to very complex ones, the main determinant of the structure being "only" the rates of mutations and rearrangements. Of course, this does not imply that organisms sharing the same mutation/rearrangement rates will have a similar structure regardless of the complexity of their respective environments. However, we can deduce from our results that the molecular complexity of the organism will be bound by robustness constraints, meaning that the mutation/rearrangement rate will remain a major factor in determining organismal complexity.

The second set of experiments we presented in this chapter was conducted with a more demanding environment, where regulation was indeed needed for an organism to be well adapted. Considering the individuals having evolved with the same rates of mutations and rearrangements from either sets of experiments, we observed that their respective regulation networks had comparable characteristics, the number of genes, TFs and pure TFs being of the same order. Even though the number of experiments we have at our disposal to date is far from sufficient, these observations are very interesting, suggesting that the complexity of regulation networks is strongly determined by indirect selective pressures and that this pressure can be high enough to over-rule the pressure due to the complexity of the environment. The improvements we have brought to the implementation of the model will allow us to test this hypothesis more thoroughly in the near future. We are looking forward to running full scale experiments using the R-Aevol model in more or less demanding environments and with a wide range of mutation and rearrangement rates.

Interestingly, using an *ad-hoc* data-mining process, we were able to decipher the behaviour of the complex regulation network that evolved in the model. We discovered two groups of genes that, after a simple complimentary analysis, provided us with a good insight into the general functioning of the network (figure III.17). The structures we discovered thanks to our mining process show the great potential of the use of computational evolution for benchmarking purposes. Indeed, the computational evolutionary process is able to surprise us by creating unexpected, yet efficient, structures when this can hardly be expected from

traditional benchmarking tools based upon random graph generators (Mendes et al., 2003) that cannot account for the evolutionary origins of real networks.
Overall, the R-Aevol model has great potential, both to be used as an *in silico* experimental evolution platform to further understand the evolutionary dynamics of regulation networks and to produce various datasets that will enable the scientific community to test *e.g.* data-mining, phylogeny reconstruction or network inference algorithms. Besides, since R-Aevol allows us to generate networks of variable sizes, we can produce collections of benchmarks of variable complexity. We are confident that both directions can help practitioners to better take up the challenges of evolutionary systems biology.

Chapter IV

Homology-Driven Recombination in Aevol

The results presented in this chapter have been published in Parsons et al. (2011).

1 Introduction

In the previous chapters, we have discussed causes and consequences of the second-order selection of a specific level of mutational variability of the phenotype. We showed, in particular, that this indirect pressure is exerted among other things on non-coding sequences because of chromosomal rearrangements. Indeed, non-coding sequences provide an additional substrate for rearrangement breakpoints, thereby increasing the overall rate of rearrangements (which can impact genes even when all the breakpoints are in non-coding sequences).
In the original Aevol model, rearrangement breakpoints were randomly chosen. This modelling choice was due to the great computational cost involved in alignment search. In real organisms however, rearrangements occur preferentially at breakpoints that are similar in sequence. For that matter, and because rearrangements are at the centre of the second-order pressures we want to study, Carole Knibbe stated in the conclusion of her PhD thesis, that "a finer modelling of rearrangement mechanisms constitutes a priority" – translated from French – (Knibbe, 2006). Indeed rearrangements are very powerful genetic variation mechanisms that can hence have dramatic effects on the genome of the organisms, which means they are very dangerous.
Rearrangements whose breakpoints are similar in sequence are called *homologous* rearrangements. By contrast, we call here *nonhomologous* rearrangements those occurring between sequences of low similarity. It is tempting to think that, because homologous rearrangements are partially directed, they could be less dangerous than rearrangements occurring at random points.

To investigate the role of homologous rearrangements in genome evolution, we extended the Aevol model to introduce a sensitivity to sequence similarity in the rearrangement process: in this extended model, a rearrangement is more likely to occur between similar sequences (homologous recombination) but remains possible, although at a low probability, when the breakpoints differ (nonhomologous recombination).
Remarkably, this model of homology-driven recombination can also be used to model horizontal transfer, where DNA is transferred between different organisms. In this context, the sensitivity to sequence similarity should allow for allelic recombination, which should in turn provide a way for evolution to circumvent the problems of linkage disequilibrium (hitchhiking), that is known to play a major role in indirect selection phenomena (Sniegowski et al., 2000).
In this chapter, I will first present an overview of the main mechanisms responsible for chromosomal rearrangements in prokaryotes. Then, I will introduce the alignment search algorithm I have developed and integrated into the Aevol model. Finally, I will present a set of experiments that allowed us to validate the model and to uncover an intricate relation between local mutations, homologous rearrangements and nonhomologous rearrangements. As for questions regarding horizontal transfer, they will be addressed in the next chapter.

2 Chromosomal Rearrangements in Prokaryotes

The two-complementary-strands structure of DNA is essential for most genetic mechanisms, in particular that of DNA replication that is achieved through the separation of the two strands which are subsequently complemented by newly synthesized strands. This process of DNA replication is of an incredible complexity and, even though the dedicated cellular machinery is very sophisticated, the synthesized sequences are not altogether free of errors. Point mutation and indels are patent examples of these errors, and the fact that the very enzyme responsible for the synthesis of the new strands (DNA polymerase III), itself contains an error-correction (proofreading) unit (Lewin, 2007) reflects their frequency. But point mutations and indels are not the only mistakes that DNA polymerases can make. When a DNA segment with a sequence similar to that being replicated is close to the replication fork (where the replication takes place), the polymerase can actually "jump" from one strand to the other and go on replicating the wrong sequence. Hence, the sequence thus produced will be a kind of patchwork of different sequences, or in other words, a rearranged sequence. As shown in figure IV.1, this kind of error can lead to large duplications or large deletions (Higgins, 2005; Lewin, 2007).
Rearrangements can also be a side-effect of error-correction mechanisms. Being the support for the genetic information, DNA is very precious for the organism. It is nonetheless subject to a lot of stress and continuously undergoes alterations of different forms. To counter these alterations, various error-correction mechanisms have evolved that can for instance replace a mismatched base by the correct one, re-synthesize a missing part of a DNA strand using the complementary strand as a template or even rejoin sequences that have been cut on both strands simultaneously (Double Stand Break, DSB). It is the latter case that interests us most because it can lead to a rearrangement. A DSB is an event during which a DNA duplex is completely cut, yielding two independent free-ended DNA

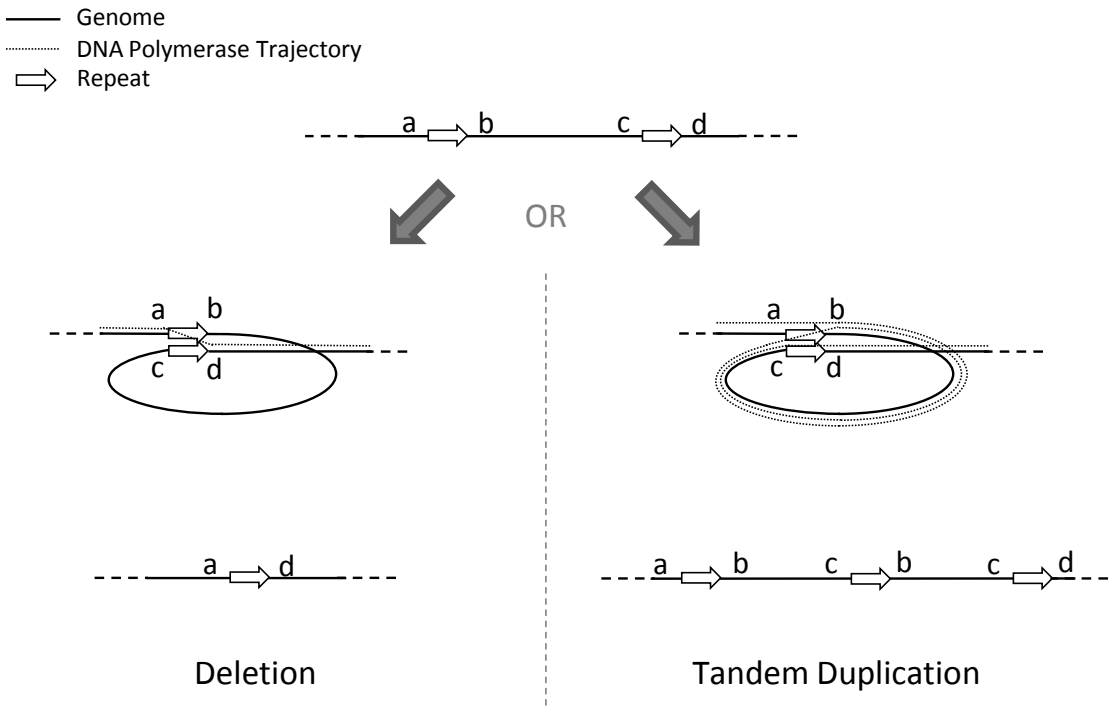


Figure IV.1 – During replication, the DNA polymerase sometimes make errors, "jumping" from one strand to the next. This can lead to large tandem duplications or large deletions.

duplexes. When a DSB occurs, the cell *must* restore the integrity of its genetic material or else it dies. Fortunately, DNA replication being continuously performed, most of the DNA is present in several copies in the cell. It is thence possible to repair the damaged sequence using one of its copies as a template. When a DSB is detected, exonucleases hydrolyse the extremity of one of the damaged strands (the recipient), thus exposing a segment of single-stranded DNA upon which RecA proteins will then bind. These proteins will allow the single-stranded segment to invade a DNA duplex (the donor) whose sequence is similar to that of the recipient, thus forming a D-loop. This D-loop will then be enlarged through an elongation process similar to that of the main DNA replication process, that actually synthesizes the segment that was missing on the former duplex. Once all the missing segment has been re-synthesized, the displaced strand migrates to its former duplex and the missing strand is re-synthesized. This finally results in two duplexes attached by two Holliday junctions (one at each extremity of the re-synthesized sequence). Depending on the resolution of these Holliday junctions (through which sequences stay linked together) and on whether the donor was indeed a copy of the missing fragment or just a similar sequence, the result can be either a correctly repaired DNA duplex or, more likely, a pair of recombinant DNA.

There are other mechanisms that can cause chromosomal rearrangements, in particular site-specific rearrangements that can be caused by transposable elements and in particular insertion sequences (IS) that are known to favour rearrangements in prokaryotes (Rohmer et al., 2007). Insertion sequences are small DNA structures flanked by inverted repeats, which bear a gene coding for transposase, an enzyme that allows the IS to be excised

and reinserted elsewhere on the genome. The transposition process can sometimes cause rearrangements as a side-effect, when two transposable elements are involved in the same event, the sequence in between being excised along with the transposable element itself.

2.1 Modelling Rearrangements

Chromosomal rearrangements are hence the consequence of very complex mechanisms which, were they directly implemented in an evolutionary model, would consistently render it intractable and unusable. However, despite the diversity of biochemical mechanisms that can lead to chromosomal rearrangements, the conditions necessary to their occurrence as well as their effects on a chromosomal scale, are very similar. Indeed, most rearrangements are the consequence of errors, themselves being facilitated by the presence of repeated sequences or similar sequences along the genome. The effects are also similar: modifications in the number and the order of subsequences of the genome, that can be either duplications, deletions, translocations or inversions (figures IV.1, IV.2and IV.3)[1].

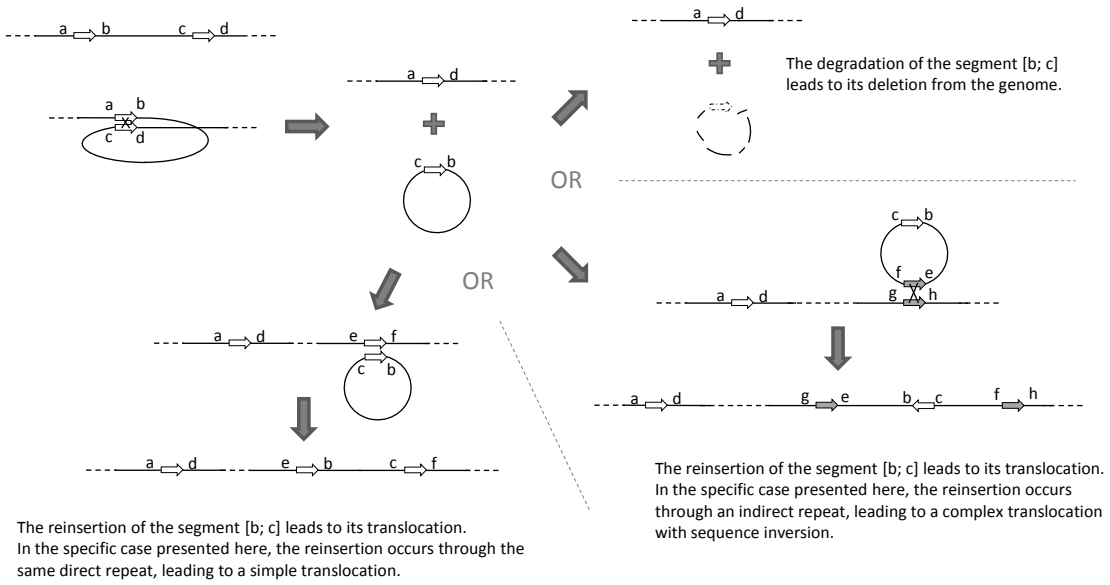


Figure IV.2 – Chromosomal rearrangements can occur by means of some error-repair mechanisms. A direct repeat in the genome sequence can lead to the excision of a segment of the genome. Then, the excised segment can be either hydrolysed (producing a deletion – **top-right**) or reinserted into the genome, leading to a translocation. If the segment is reinserted using the same breakpoint as for its excision, the whole process will produce a simple translocation, where the translocated segment is simply displaced (**bottom-left**). If the breakpoint involved in the reinsertion of the segment is different from the one involved in its excision, the order of the sequences will not be maintained, producing a "complex translocation" (**bottom-right**). Note that in both cases, if the second alignment involves indirect repeats, the translocated segment will also be inverted.

1. Here, we modelled only *in situ* (tandem) duplications. Indeed, inserting the duplicated sequence elsewhere would require another alignment to be found.

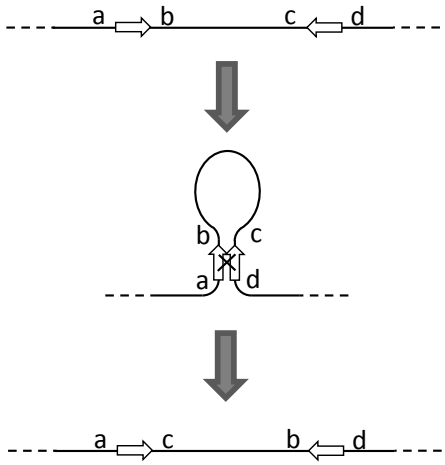


Figure IV.3 – An indirect repeat in the genome sequence can trigger a large inversion through error-repair mechanisms.

From there, we can model the rearrangement process at a mesoscopic level by considering only these key principles, common throughout the different mechanisms that can lead to a rearrangement. The granularity of the model will hence be at the level of the rearrangement itself, and the probability of a rearrangement occurring will directly depend on the degree of similarity between the potential breakpoints. We thus need, in our model, a way to detect and quantify sequence similarities, in other words, we need to define an algorithm that searches for homologies within the sequences.
In previous versions of the Aevol model, this sensitivity to sequences similarity had to be abandoned because of the computational cost it implies.

3 Searching for homologies

Searching for homologies between sequences (mostly proteic or nucleic) is a central question in the field of bioinformatics. This highly combinatorial problem is well known to be very difficult and, for over 40 years, a great deal of work has been dedicated to finding algorithms to solve it (Needleman and Wunsch, 1970; Smith, 1981; Wilbur and Lipman, 1983; Lipman and Pearson, 1985; Altschul et al., 1990).
The degree of similarity between two sequences can be quantified thanks to a scoring function, similar to an edit distance: the alignment score between two sequences will hence depend on the number and the type of mutations (point mutations or indels) that are necessary to switch from one sequence to the other. A *substitution matrix* is used to determine the reward or penalty corresponding to each pair of residues [1] when placed in vis-à-vis, and indels are accounted for by *gaps* in the alignments: one or more residues from one sequence can be associated with nothing on the other sequences. This accounts for either a small insertion on the first sequence or a small deletion on the second. The

1. The term "residue" refers to the unitary elements which the sequences are made up of. Nucleotides and amino-acids are the respective residues of nucleic and proteic sequences.

contribution of these gaps to the alignment score is usually composed of both a fixed cost for opening the gap and a cost for its extension that depends of its length.
The result of an alignment search is often presented in the form of a *dot plot*, where two compared sequences are plotted with respect to each other. In this kind of visualization, alignments between the two sequences take the form of diagonals or collections of diagonals linked together when *gaps* are allowed (figure IV.4).

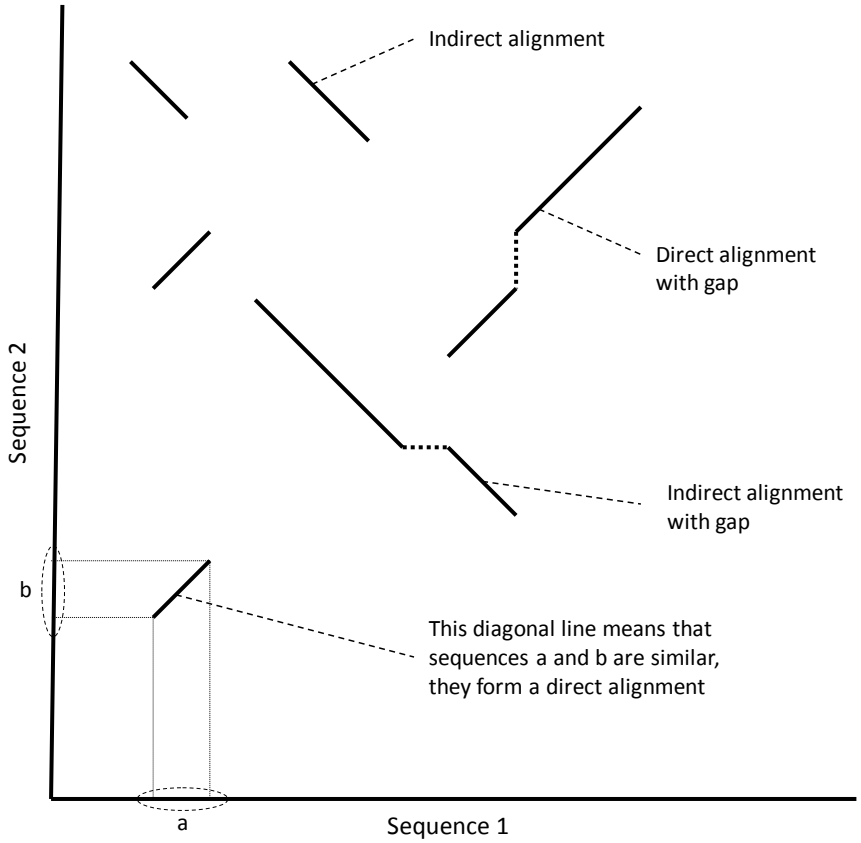


Figure IV.4 – In the dot plot representation, the two compared sequences are plotted with respect to each other. Then, local similarities between the sequences take the form of diagonals, from bottom left to top right for direct alignments and from top left to bottom right for indirect alignments. Examples of both gapped and ungapped alignments are shown on the figure.

3.1 Alignment Search, a (very) Brief Chronology

In 1970, Needleman and Wunsch proposed an algorithm based on *dynamic programming* to perform a global alignment search(Needleman and Wunsch, 1970), *i.e.* that tries to align every single residue from the sequences to be aligned. However, global alignments are ill-suited for sequences of weak or partial similarity for which local alignment searches have to be performed. It was only in 1981 that Smith and Waterman proposed their famous

algorithm for the local alignment search problem (Smith, 1981). This algorithm is also based on dynamic programming. Both the Needleman-Wunsch algorithm and the Smith-Waterman algorithm perform a complete search. Their time complexity is of $O(mn)$.
Two years later, in 1983, Wilbur and Lipman presented a heuristic method (Wilbur and Lipman, 1983) that was subsequently improved and became known as FASTP (Lipman and Pearson, 1985) and FASTA (Pearson and Lipman, 1988). The idea is to first look for "promising" subsets of the search space and then proceed to a near-complete search within these narrower search spaces. In 1990, Altschul et al. pushed this idea further still and came up with their Basic Local Alignment Search Tool – BLAST – (Altschul et al., 1990), that was to be further improved, in particular with Gapped BLAST and PSI-BLAST (Altschul et al., 1997). Even though FASTA and BLAST proceed in a slightly different way, the methods for identifying promising regions in the search space are similar: the basic idea is that a biologically significant alignment will contain small sequences that match exactly or almost exactly. These small but strong alignments are respectively referred to as *k-tuples* in FASTA and *hits* in BLAST; we will use the term *hit* for the remainder of this thesis. Once all the hits have been found, they can be extended, meaning that the sequences surrounding the hits are looked for in a attempt to further increase the alignment score, *i.e.* to verify whether they belong to a biologically significant alignment.

3.2 Searching for Alignments in the Context of Digital Genetics

BLAST and its extensions are very efficient local alignment search tools. Today, BLAST is probably the most widely used alignment search tool in the bioinformatics community. However, even though it is very efficient, "blasting" sequences remains nonetheless relatively long (a few seconds). In the specific case of Aevol, the alignment search algorithm has to be integrated into the evolutionary algorithm which means that it would have to be performed at each replication to find the candidate breakpoints for rearrangements. Now if a single BLAST of a whole genome on itself is not computationally prohibitive, a systematic search before each replication would take a virtually endless computation time [1], which makes it an ill-suited solution. Thus, we must consider other possibilities that take advantage of the specificities of *in silico* evolution [2].
Since in Digital Genetics, genomes are algorithmically built (and not observed), one could choose to represent them using an efficient data structure. For example, structuring the genome as a suffix-tree could have been a solution to reduce the computational time of alignment searches. However, given the size of the genomes in Aevol (from a few hundreds to a million bp), this would consistently have led to a memory size explosion. Besides, there is no guarantee that this solution would have produced a substantial improvement. As a matter of fact, the idea of taking sequence similarity into account in the rearrangement process had been abandoned within the scope of Carole Knibbe's PhD because of the forbidding computational cost it represented.

1. As an example, assuming that "blasting" a genome on itself takes only 1 second, the overall time spent on alignment searches during a single typical Aevol run of population size 1,000 during 50,000 generations, would be of over one year and a half.

2. Homologous cross-over operators have been proposed in the context of Genetic Programming (D'haeseleer, 1994; Poli and Langdon, 1997; Nordin et al., 1999; Langdon, 2000; Defoin Platel et al., 2003; Defoin Platel, 2004). However these works aimed at optimization goals while we are interested in studying the process itself.

However, we can take advantage of another specific characteristics of digital genetics: digital genetics frameworks are simulation frameworks, not data-analysis ones. Indeed, all the algorithms mentioned above were designed to find *all* the biologically significant alignments between two sequences. In our specific case however, we do not need such an exhaustive search; rather, we *only* need to find a few correct alignments to perform a rearrangement. Depending on the type of the rearrangement considered (duplication, deletion, inversion or translocation), only one or two alignments are needed. Moreover, we want most rearrangements to occur between highly similar sequences so our need for sensitivity is quite low. Thus our problematic is rather to find *a few* high-quality alignments, and not *all* the moderately similar pairs of sequences. If this new formulation does not solve our algorithmic question, it nonetheless allows us to widen the range of methods available, among which some might prove better adapted to this particular case.

3.3 Intermittent Search Strategies

It is surprising that, while searching for sequence alignments is algorithmically difficult, alignments are nonetheless a central feature of many cellular processes. It would thus not be too surprising to identify efficient solutions for finding alignments within these particular biological processes. It was shown for instance, that repressors (proteins that can bind to promoter operator sites and thereby reduce the transcription rate) can localize their target site way faster than would be possible in the context of either a three-dimensional diffusion in the medium or a unidimensional diffusion along the DNA strands(Riggs et al., 1970). A few years later, Richter and Eigen (1974) proposed that this high speed association rate is due to the "unspecific binding of repressor to nonoperator DNA with subsequent diffusion along the chain": the repressor, they suggested, alternated between stages of three-dimensional diffusion in the medium and stages of one-dimensional diffusion along the DNA molecule.
Search strategies such as this one, alternating between stages of intensive local search and stages of fast, blind, movements between different regions of the search space are referred to as intermittent search strategies (Bénichou et al., 2011). Intermittent search strategies are widely observed at different scales in natural systems, particularly in animal and human foraging behaviours (Viswanathan et al., 1996; Bénichou et al., 2005; Shlesinger, 2006; Brown et al., 2007; Edwards et al., 2007) or in target localization by chemicals (von Hippel and Berg, 1989; Coppey et al., 2004; Bénichou et al., 2006). This kind of search is indeed very efficient and has been shown to be optimal in the case of persisting targets (Viswanathan et al., 1999). So we based our alignment search algorithm on the principle of intermittent search, using BLAST-inspired methods for local search stages.

4 An Algorithmic Model of Intermittent Alignment Search

As we have previously stated, biochemical mechanisms that can lead to chromosomal rearrangements are both complex and numerous. However, the modelling choices that were presented in section 2.1 allow us to simplify these mechanisms while retaining the biological plausibility of the model. Thus, the starting point of our model, the atomic event

we will consider, is the identification of an alignment between two DNA sequences. The search for these alignments takes place within the evolution of a population of artificial organisms. At each replication, the genome to be replicated will thus be searched for both direct and indirect alignments to find the candidate breakpoints for rearrangements. Direct alignments can lead to either duplications, deletions or translocations while indirect alignments can only lead to inversions [1]. This search will be conducted locally around points assumed to be located in the same three-dimensional neighbourhood.

At any time during the life of a cell, its chromosome has a given spatial conformation. A simple way of modelling this spatial conformation is to consider local neighbourhoods. We can hence consider that there is a set of pairs of DNA segments that are physically close together. The conformation of the chromosome being a dense supercoil, the three-dimensional proximity of any two segments of the chromosome does not depend on their relative positions on the sequence: in a first approximation, two diametrically opposed segments on the sequence can be considered as likely to be neighbours in the three-dimensional conformation than two segments that are close together on the sequence. This independence allows us to model this phenomenon in a very simple way, namely in the form of a random drawing of pairs of *neighbouring points*, following a uniform distribution on the size of the chromosome. The number of pairs of points to be drawn will depend on how densely packed the genome is, *i.e.* on its degree of supercoiling. We will characterize this degree of supercoiling by a degree of vicinity, referred to as the *neighbourhood rate* μ_n that we will assume to be constant over time and throughout the entire genome. The number of pairs of points to be drawn will then be proportional to the genome length: $nb\_pairs = L.\mu_n$.

This simple process enables us to implement the global search phase of our intermittent search strategy (algorithm 1). Then, for each candidate pair of points, a local alignments search will be performed to determine the existence of similarities between the surrounding sequences, either in a direct or indirect sense. Figure IV.5 illustrates the intermittent search strategy.

The surroundings of a candidate pair of points, *i.e.* the local search space, is presented in figures IV.6 and IV.7. It is assumed that, for each pair of *neighbouring points*, the sequences within a given distance defined by the parameter $half\_length$ are face to face. These sequences will be referred to as the *working zone* for the local search. It is also considered that these two sequences can slip with regard to each other, within a certain threshold fixed by the max_shift parameter. Then, each nucleotide that falls within the working zone of one sequence will be tested versus its direct vis-à-vis as well as its max_shift neighbours both upstream and downstream. This will produce an extension on each side of the working zone to avoid border effects and guarantee that each nucleotide that belongs to the working zone is tested versus the same number of vis-à-vis.

At the global level, this search process is hence the combination of a Monte Carlo process, randomly drawing pairs of points within the genome, and of a local search process between the zones surrounding each of these points. This global process is hence indeed an intermittent search as we had previously proposed. However, although this type of search seems well suited for our problem, the local search process is still problematic, even though the search space will be small and of constant size and shape. We propose

1. Indirect alignments can also play a role in the reinsertion of a segment during a translocation as shown figure IV.2

$initial\_nb\_pairs \leftarrow L * \mu_n$
$nb\_pairs \leftarrow initial\_nb\_pairs$
while $nb\_pairs > 0$ **do**
  Draw 2 random positions $pos1$ and $pos2$
  Draw type of rearrangement
  if *Inversion* **then** $sense \leftarrow indirect$
  else $sense \leftarrow direct$
  Draw minimal alignment score for a rearrangement to occur
  Search Alignment($pos1$, $pos2$, $sense$, min_score)
  if *Alignment found* **then**
    Proceed to Rearrangement
    Update L
  end
  $nb\_pairs \leftarrow nb\_pairs - 1$
  $nb\_pairs \leftarrow \frac{nb\_pairs}{initial\_nb\_pairs} * L * \mu_n$
end

Algorithm 1: Aevol Rearrangement Process Algorithm

here two local alignment search strategies, in which one allows for gaps while the other does not.

4.1 Local Alignment Search Allowing for Gaps

The design of our local search algorithm allowing for gaps is based upon the principles proposed in BLAST (Altschul et al., 1990) and more specifically in its extensions PSI-BLAST and Gapped BLAST (Altschul et al., 1997). BLAST uses a heuristic that considers that any biologically significant alignment is very likely to contain small sequences that align almost perfectly. It proceeds in two steps, the first consisting in a complete search for these small but highly similar sequences (*hits*), and the second, in an attempt to improve the score of these *hits* by extending them, which actually corresponds to checking whether they belong to a significative alignment.
The PSI-BLAST extension (Altschul et al., 1997) follows the idea that because the *hit* extension stage is the most computationally costly, the number of *hits* to be extended must be limited as much as possible while not sacrificing the sensitivity of the whole search process. In this perspective, it was proposed to consider pairs of *hits*, namely *two-hits*, instead of single *hits*, a *two-hit* being made up of two *hits* that are on the same diagonal within a predefined distance. Obviously, to preserve a similar level of sensitivity while considering *hits* by pairs, the size of single hits must substantially be reduced, which implies that single *hits* will be a lot more numerous. However, the number of *two-hits*, and hence of extensions to be performed, will be substantially lower using this solution than with the original algorithm. Hence, considering the respective time consumption of *hit* generation and extension, resorting to *two-hits* proves to be a lot more efficient.
The original BLAST algorithm does not consider gapped alignments explicitly; rather, it considers sets of ungapped alignments and computes a statistical estimation of the combined alignment. The Gapped BLAST extension was designed to directly consider

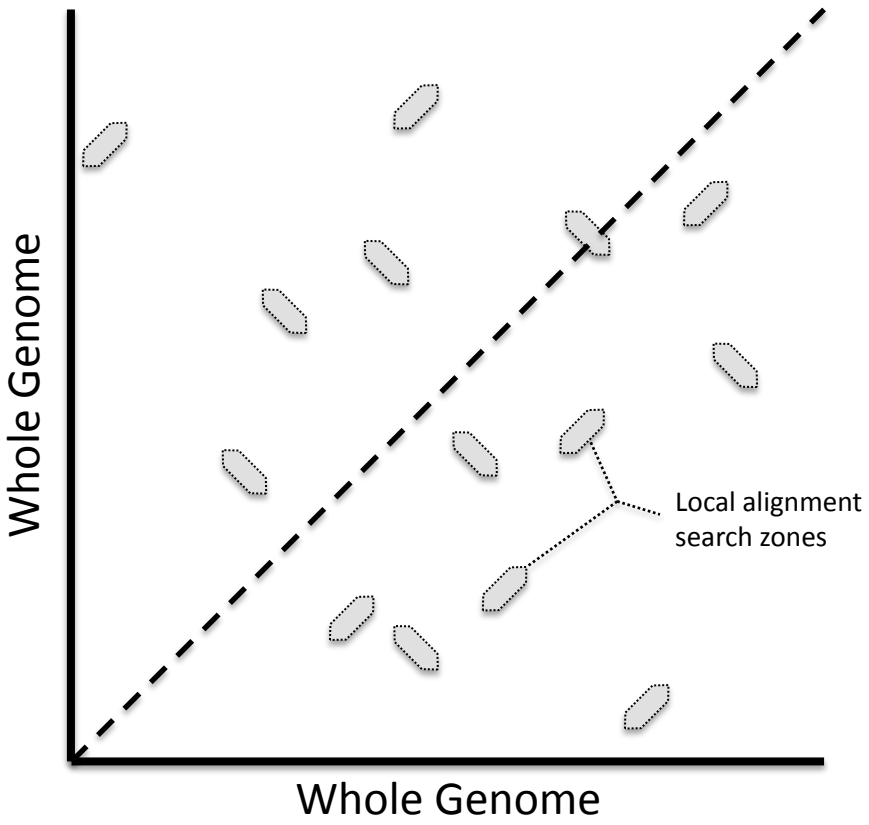
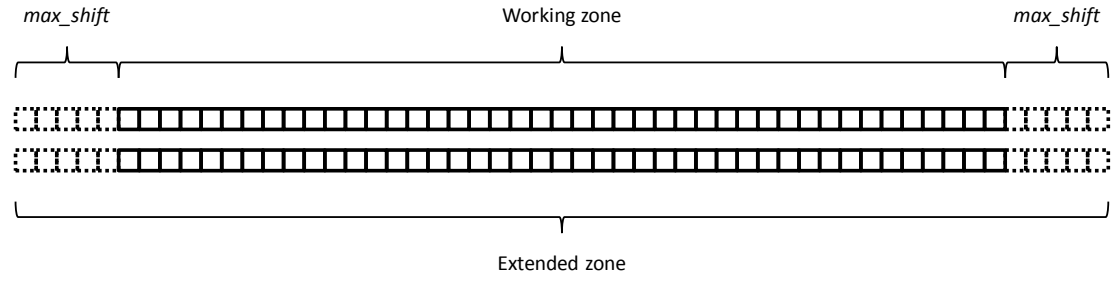


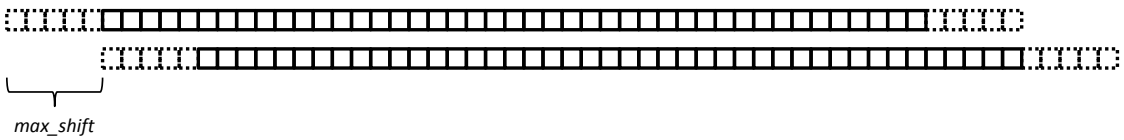
Figure IV.5 – For each candidate pair of points ($nb\_pairs = L.\mu_n$ with L, the size of the genome and μ_n, the neighbourhood rate), a local search is performed to determine the existence of an alignment of sufficient score.

gaps in the local search process. It was considered that, given the cost of opening a gap in an alignment, it is only necessary to consider this eventuality for very promising, (*i.e.* high score) ungapped alignments. Thus, although the computational time needed for considering every possible gapped extension is prohibitive (it corresponds to a variant of the Smith-Waterman algorithm), this exploration is seldom performed, yielding only a mild additional cost to the overall search process.

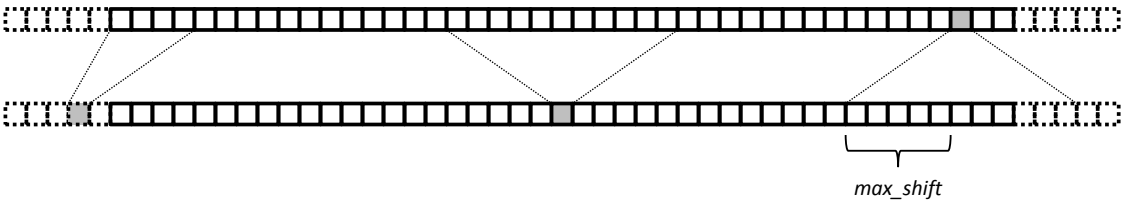
Our algorithm follows the same general search scheme, first looking for *hits* and *two-hits*, and then extending them. In our algorithm, the *hit* generation step is a complete search of all the pairs of sequences of a given size that have a score greater than a given threshold. *Hits* are looked for within the local search space (see section 4). Once all the *hits* have been generated, they are confronted with one another to form *two-hits*, following the idea of the PSI-BALST algorithm. However, this idea, combined with that of Gapped BLAST, was pushed one step further: in our specific case, most rearrangements require a very high score alignment to occur. Furthermore, the search process being included in the evolutionary loop, it must be very computationally efficient. For both these reasons, we authorized at most one gap per alignment, this gap being opened only when it allows two *two-hits* from different diagonals to be joined. Using this method, the calculation times



(a) Working zone and extensions.



(b) Sequences are allowed to slide on each other.



(c) Examples of vis-à-vis to be tested.

Figure IV.6 – **(a)**: The local search space is defined as a working zone and two extensions on each sequence. The working zone consists of the $half\_size$ nucleotides surrounding the original point in both direction and the extensions, of the subsequent max_shift nucleotides ($half\_size$ and max_shift are parameters of the model). **(b)(c)**: The sequences are allowed to slide with respect to each other during the local search process. Thus, each nucleotide within the working zone of a sequence will be confronted to $2 * max\_shift + 1$ nucleotides of the other sequence, either from its working zone or from one of its extensions.

are a lot shorter. Indeed, exploring gapped alignments on the basis of two *two-hits* has a linear time complexity. Besides, these explorations are seldom performed because they require two *two-hits* to be found within a short distance in the search space (though on different diagonals). The *multiple-hits* (either *two-hits* or pairs of *two-hits* linked by a gap) are then extended to try and improve their score.

4.2 Ungapped Local Alignment Search

The ungapped local search algorithm is a lot simpler than the gapped one. Indeed, when gaps are not allowed, the effective search space is substantially reduced. In fact, since the local search space is rather small, it is reduced to a point where heuristics are not needed any more and a complete search algorithm can be used. Each diagonal in the search space is parsed linearly, each match between two nucleotides being rewarded one point and each mismatch resulting in a penalty of two points.

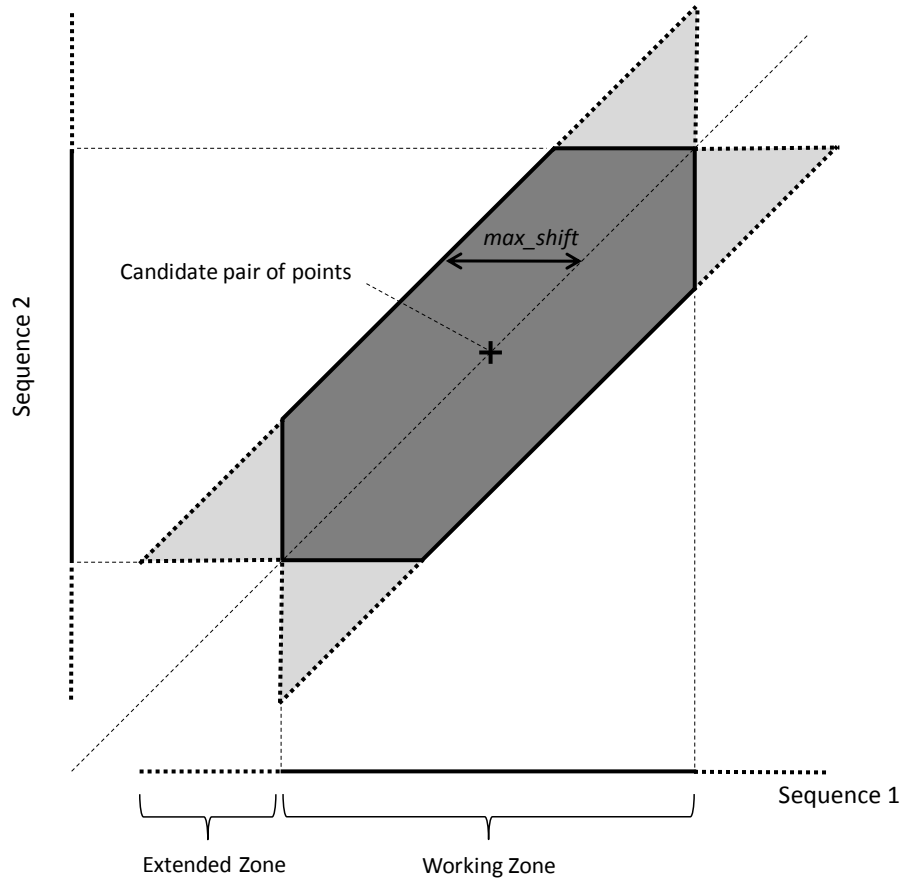


Figure IV.7 – The local search space seen in a dot plot fashion. The dark grey zone corresponds to the working zone while light grey corresponds to the extension zones

4.3 Algorithm Improvements

Searching for alignments, either allowing for gaps or not, is computationally costly. A great deal of effort was hence dedicated to optimizing the algorithm and the implementation of both the overall search process and the local search. Most of these optimizations are out of the scope of this dissertation, being purely technical and having no impact on the model. However, one particular improvement of the algorithm does require our attention. The probability of a rearrangement taking place between two given breakpoints depends on the score of the alignment between the sequences around these points. However, in biology, not all the similar sequences that happen to be close together in the nucleoplasm will produce a rearrangement. From the point of view of the program, once the best alignment in the local search space has been found, the corresponding score has been computed, and hence p_{rear} is known, a rearrangement will occur if $rand < p_{rear}$ ($rand$ being randomly drawn following a uniform distribution in $[0, 1[$). Note that this enables to allow nonhomologous rearrangements to occur in a controlled proportion. Figure IV.8 shows probability of finding an alignment of a given score on a random sequence as well as the function $p_{rear}(score)$ we used in these experiments. This particular function (see equation IV.1) with $\alpha = 50$ and $\lambda = 4$ yields a reasonable trade-off between homologous

and nonhomologous rearrangements.

$$p_{rear}(score) = \frac{1}{1 + \exp -\frac{score - \alpha}{\lambda}} \tag{IV.1}$$

Obviously, whether *rand* is drawn before or after the alignment search will not change anything. However, knowing its value before the search, we can compute the minimum score needed for a rearrangement to occur in this particular context. It is hence possible to direct the local search process and hence to speed it up: knowing the minimum needed score, it is possible, instead of searching for the best alignment in the local search space, to look for the first alignment with a score at least equal to this minimum. Besides, when the required score is high, heuristics can be used to discard whole regions of the search space that are not promising enough, *i.e.* when they don't seem likely to contain such a high score alignment. Using the first "good enough" alignment instead of the best one in the local search space is a modification of the model. In fact, it probably is an improvement of the model since high score alignments that are not locally optimal are no longer silenced by the local optimum.

4.4 Performance Tests

The main goal of the work presented in this chapter was to design an alignment search method fast enough to be integrated into an evolution cycle such as that used in Aevol. Now, although on the global scale, the proposed intermittent search process allows a significant reduction in the computational cost of alignment search, it still requires the local search process to be very efficient. Indeed, this local search will be performed many times for each replication, in the context of the evolution of a whole population of individuals for thousands of generations. This local search being repeated up to trillions of times during a single simulation, any improvement of this search will have a great impact on the overall computational time needed.
The complexity of a complete search for alignments between two sequences is such that it is impossible to measure the execution times of such an alignment search between long sequences such as whole genomes. For this reason, we ran the performance tests specifically on the local search process, which represents the core of our algorithm. In order to study the performance of the local search algorithms we have proposed, we developed a benchmark algorithm that performs a *complete* search in the same search space as the other algorithms. It is important to note that, although this benchmark indeed performs a *complete* search, it has nonetheless been greatly optimized and is itself a lot faster than a brute force algorithm would be.
Figure IV.9 shows the average computation time of one local search for an alignment with a score of at least 50 between random sequences, as a function of the size of the search space (determined by the half_length of the working zone) for all three algorithms (the complete algorithm and the algorithms we propose, either allowing for gaps or not). Note that since we used random sequences, alignments are seldom found, meaning that in most cases, the entire local search space is covered. These measures hence correspond to worst-case values. As expected, the measured computation times for each of these algorithms are very different: the intermittent search strategy allowing for gaps runs 15-fold faster

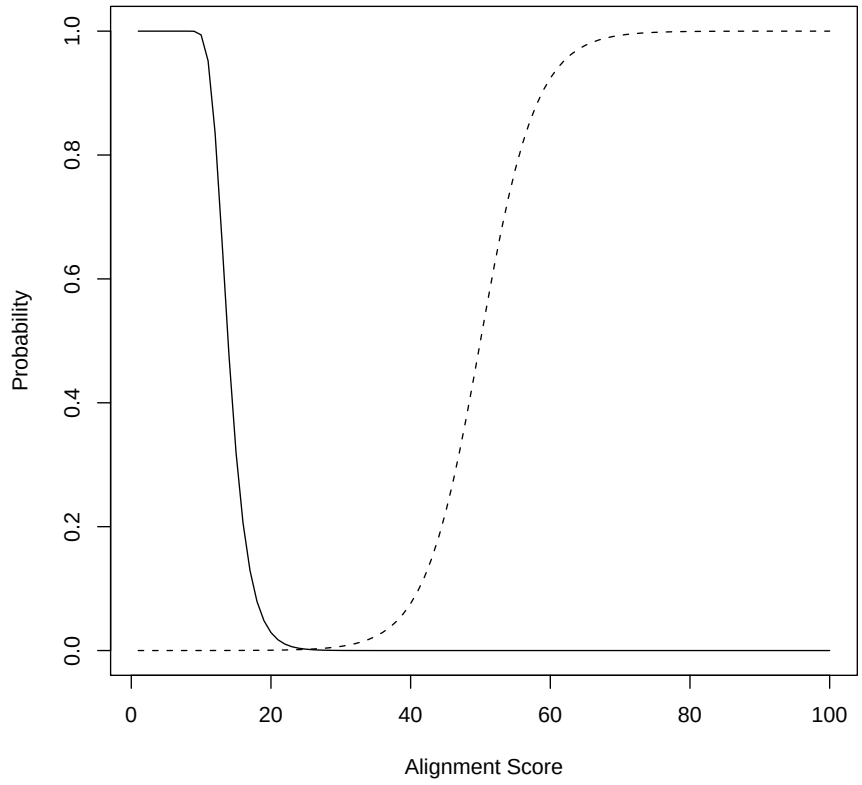


Figure IV.8 – Probability of finding an alignment of the corresponding score on a random sequence (solid line) and probability of an alignment of the corresponding score to produce a rearrangement (dashed line). Preliminary experiments allowed us to calibrate the function we used to obtain a reasonable trade-off between homologous and nonhomologous rearrangements.

than the complete search and the intermittent search without gaps, another 2.8-fold faster than that allowing for gaps.
Let us consider these computation times in the context of a $50,000$ generation long simulation in Aevol. Preliminary tests showed us that, using the function shown in figure IV.8 to map alignment scores with rearrangement probabilities within a search space defined by $half\_length = 50$ and $max\_shift = 20$, for a neighbourhood rate of $\mu_n = 5 \times 10^{-2}$, the evolved genomes would have a length of approximately $10,000$ bp. We can then evaluate the overall number of local alignment searches to be performed during the whole evolutionary process as the product

$$nb\_generations \times population\_size \times genome\_size \times \mu_n = 2.5 \times 10^{11} \qquad \text{(IV.2)}$$

According to the corresponding values in our test case, simulating this evolution using the complete search would then require almost 5 years, the intermittent search strategy allowing for gaps would reduce the time needed to around four months and the intermittent

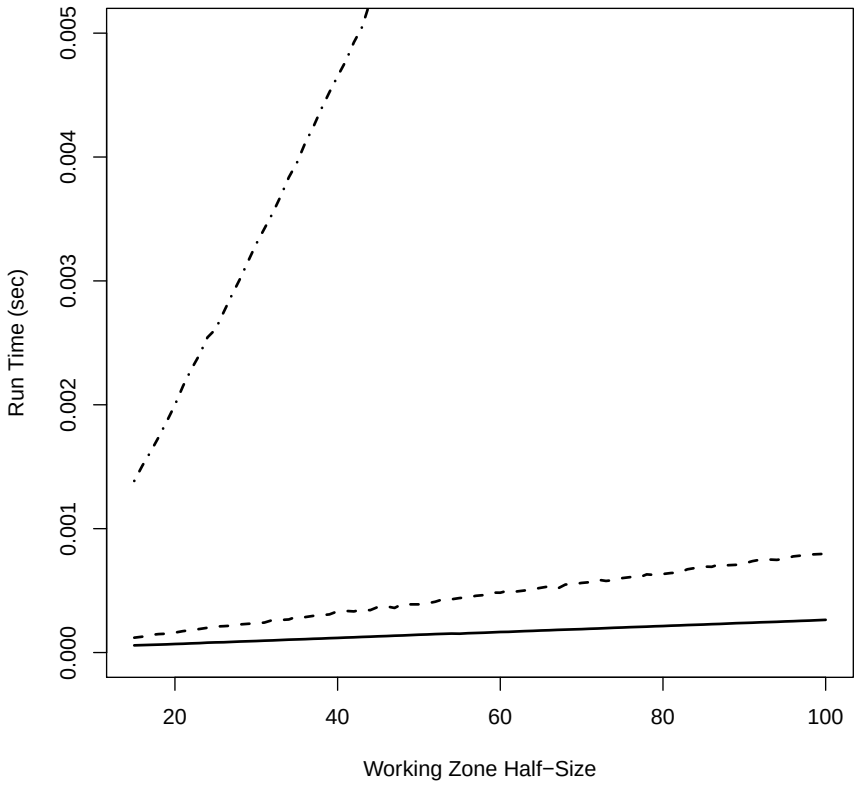


Figure IV.9 – Average execution time for a single local alignment search as a function of the working zone half-size for **solid line**: the intermittent search strategy with no gaps allowed, **dashed line**: the intermittent search strategy allowing for gaps and **mixed line**: the complete local search algorithm.

search strategy with no gaps allowed, just over a month. As we have stated, these performance tests were conducted using random sequences. These are hence worst-case values and we expect simulations to be faster. However, even though our intermittent search algorithm allowing for gaps is indeed very efficient compared to the complete search, conducting a full scale experiment using this search strategy would require way too much computational power. We will hence conduct our experiments using the simplest model that does not allow for gapped alignments.

5 Validating Our Homologous Rearrangements Model

5.1 Experimental Setup

In this chapter, we have proposed a computationally tractable model of chromosomal rearrangements that accounts for the sensitivity of rearrangement mechanisms to sequence similarity. To validate this model, we conducted a large-scale experiment of evolution

in conditions comparable to those previously tested with the Aevol model. We let 60 populations of 1,000 asexual individuals evolve during 20,000 generations in near identical conditions where the only changing parameters were the mutation rate (one common rate μ_m for the three different types of local mutations – tested values: $\mu_m = 5 \times 10^{-6}$, 10^{-5}, 5×10^{-5} and 10^{-4}) and the neighbourhood rate (μ_n, tested values: $\mu_n = 10^{-2}$, 5×10^{-2}, 10^{-1} and 5×10^{-1}). The complete set of parameters used in these experiments is presented in table IV.1. Note that in this version of the model where the rearrangement process is driven by homology, the rearrangement rate is no longer a parameter of the model. However, figure IV.10 shows us that the spontaneous rate of rearrangements μ_r that we observed was at least partly driven by the neighbourhood rate μ_n. Thus, by acting on μ_n, we are still able to efficiently drive the spontaneous rearrangement rate.

Parameter	**Value**
N	1,000
nb_gener	50,000
$init\_length$	5,000
$init\_method$	Clonal, One Good Gene
$selection\_scheme$	Exponential Ranking
c	0.998
$E = \sum_i \alpha_i G_i$	$\alpha_1 = 0.3; G_1 : \mu = 0.1; \sigma^2 = 0.02$
	$\alpha_2 = 0.3; G_2 : \mu = 0.9; \sigma^2 = 0.02$
$env\_sampling$	300
μ_{point}	$\mu_m \in \{10^{-6}, 2 \times 10^{-6}, 5 \times 10^{-6}, 10^{-5}, 2 \times 10^{-5}, 5 \times 10^{-5}, 10^{-4}\}$
μ_{s_ins}	
μ_{s_del}	
$neighbourhood\_rate$	$\mu_n \in \{10^{-2}, 5 \times 10^{-2}, 10^{-1}, 5 \times 10^{-1}\}$
$p_{rear}(score)$	$\frac{1}{1 + \exp -\frac{score - \alpha}{\lambda}}$ with $\alpha = 50$, $\lambda = 4$
$working\_zone\_half\_size$	50
max_shift	20
max_indel_size	6
W_{max}	0.01

Table IV.1 – Parameters used in all the experiments of this chapter. The mutation rate takes its values among those proposed, one common value for the three types of local mutations.

Globally, the evolutionary process is not modified: the organisms progressively acquire new genes by duplication and modify them in such a way that the whole gene repertoire fulfils the task the organisms are selected for. All the simulations proceed qualitatively in a similar way, evolving quickly in the first stage of evolution (rapid gene acquisition

mostly by duplication-divergence) then slowing down the process of gene acquisition while optimizing the sequence of existing genes and promoters.

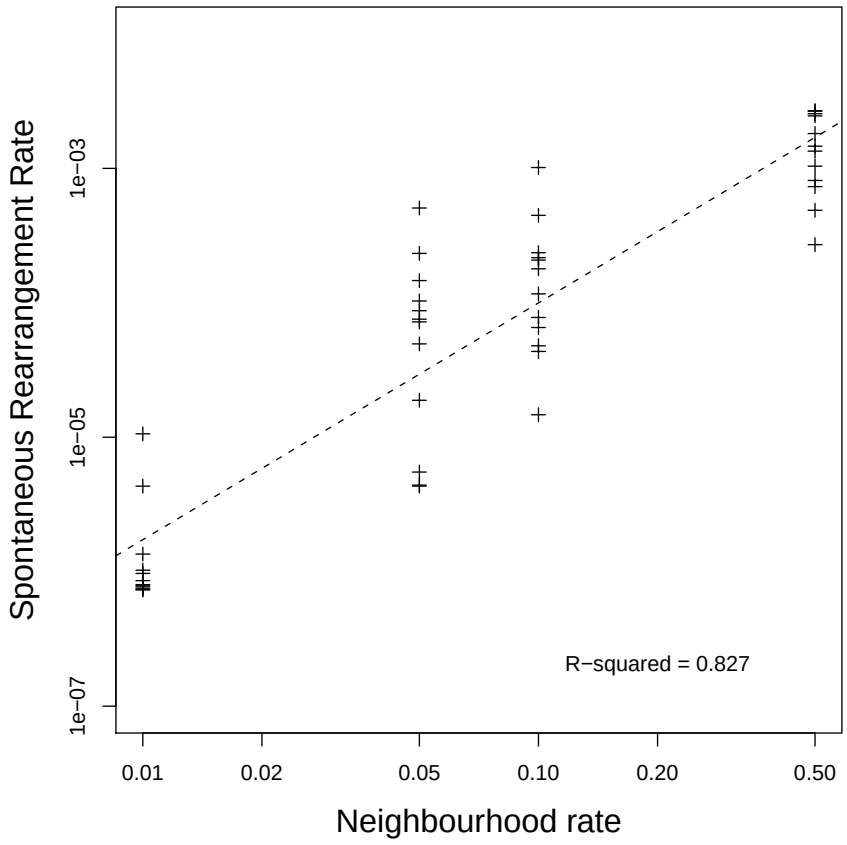
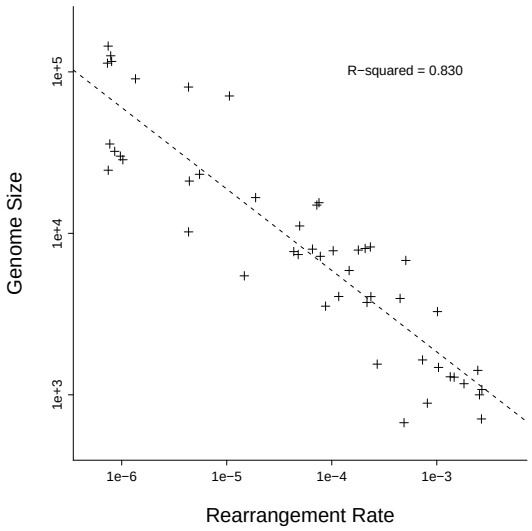
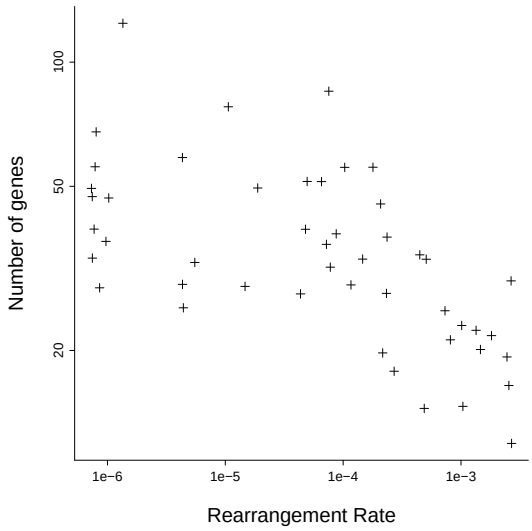


Figure IV.10 – Average spontaneous rearrangement rate observed during each simulation as a function of the (fixed) neighbourhood rate.

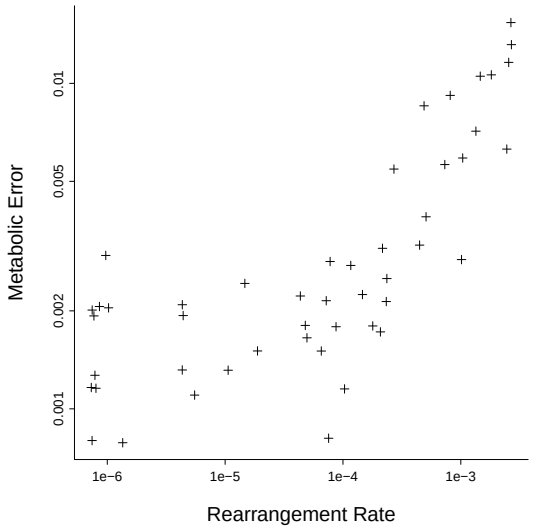
Compared to the experiments we have already presented, the rate at which rearrangements occur is no longer constant nor fixed by the experimentalist. It depends on both the neighbourhood rate μ_n and the presence of repeated sequences on the chromosome. It is hence free to evolve and could well be selected for or against. Yet, despite this added degree of freedom, the measured rearrangement rate remains a very strong determinant of genome size and content (figure IV.11). These results confirm those obtained with previous versions of the model in which the rearrangement rates were direct parameters of the model (Knibbe et al., 2007a). Even with homologous rearrangements, we find again that the spontaneous rate of rearrangement has a negative impact on fitness (figure IV.11(c)) because it still sets an upper bound on genome size and hence on the number of genes (figure IV.11(b)). However, rearrangements are also mandatory for evolution to be efficient. Indeed, an organism whose genome lost its capacity to rearrange would be unable to *e.g.* duplicate its genes. Combined to the fact that genes are very unlikely to appear *de novo*, this means that such an organism would be unable to enlarge its gene repertoire and would hardly be evolvable at all.



(a) Genome Size



(b) Number of Genes



(c) Metabolic Error

Figure IV.11 – Genome Size **(a)**, Gene Number **(b)** and Metabolic Error **(c)** of the best organism after 20,000 generations for each simulation, as a function of the spontaneous rearrangement rate.

5.2 Trade-off between homologous and nonhomologous rearrangements

Homologies are created by rearrangements (duplications), either homologous or nonhomologous. Subsequently, these repeated sequences will promote homologous rearrangements. There is hence a sort of positive feedback loop to rearrangements and this process seems to be self-maintained. However, as duplications are performed in tandem (the duplicated sequences being side-by-side), a single deletion between two repeats can cancel several

duplication events in a single step, going back from an amplified tandem array to a single copy of the sequence (Higgins, 2005). Moreover, local mutations gradually destroy the repeats created by duplication, making the overall process even more complex. There must hence be some sort of complex interactions between the mutation rate, the neighbourhood rate and the (evolving) rates of both homologous and nonhomologous rearrangements.

A rearrangement can occur between any pair of sequences, whether they are similar (homologous rearrangement) or not (nonhomologous rearrangement). However, similar sequences have a greater probability of leading to a rearrangement than sequences of low similarity. In our model, given the probability of finding alignments of different scores on a random sequence, we consider homologous those rearrangements whose breakpoints align with a score at least equal to 30. Rearrangements whose breakpoints have lower alignment scores are considered nonhomologous. The distribution of the scores of the alignments that led to rearrangements for each mutation rate and neighbourhood rate (figure IV.12) can help us understand this intricate relationship. If we consider this data vertically, we can clearly observe that the proportion of homologous rearrangements is higher when the neighbourhood rate is high. However, as we progress downwards, the distributions behave differently: while they remain nearly unchanged on the left hand side, nonhomologous rearrangements become much more frequent on the right. A noteworthy observation is that there is a great variation in the number of rearrangement events. In fact, it is not the number of nonhomologous rearrangements that increases (it actually remains stable), but rather the number of homologous rearrangements that collapses when the neighbourhood rate decreases.

The underlying phenomenon is best understood when looking at the data in a top-left to bottom-right fashion. One can then identify a phase transition between a regime of mainly homologous rearrangements at high μ_n and low μ_m, and a regime of mostly nonhomologous rearrangements at low μ_n and high μ_m. In fact, for the possibility of homologous rearrangements to be maintained along the evolutionary process, homologies must be created (by either homologous or nonhomologous duplications) at least as fast as they are destroyed by local mutations. At high neighbourhood rates, this condition is always achieved because rearrangements are numerous. However, at low neighbourhood rates, rearrangements are not so frequent and a complex interaction between chromosomal rearrangements and local mutations can appear. When mutations are very frequent, the modifications they cause in the sequence can overcome the creation of homologies and stall the whole process.

Since they correspond to the set of parameters in which these subtle interactions can occur, the four histograms at the bottom of Figure IV.12 are the most interesting. Within this line, throughout which $\mu_n = 10^{-2}$, the change in rearrangement mode from mainly nonhomologous to mainly homologous is particularly clear when the spontaneous rate of small mutations decreases.

To better understand the dynamics of homologous versus nonhomologous rearrangements, we further analysed the simulations from the left hand side, that display both the greatest proportion of homologous rearrangements (within the bottom line) and, interestingly, the best final fitness of all parameter sets. For the three runs of this parameter set ($\mu_n = 10^{-2}$ and $\mu_m = 5 \times 10^{-6}$), we kept track of the family ties during the evolution. We then retrieved the line of ancestry of the final best individual and analyzed the mutational events that occurred on this successful lineage. Except for those that occurred during the

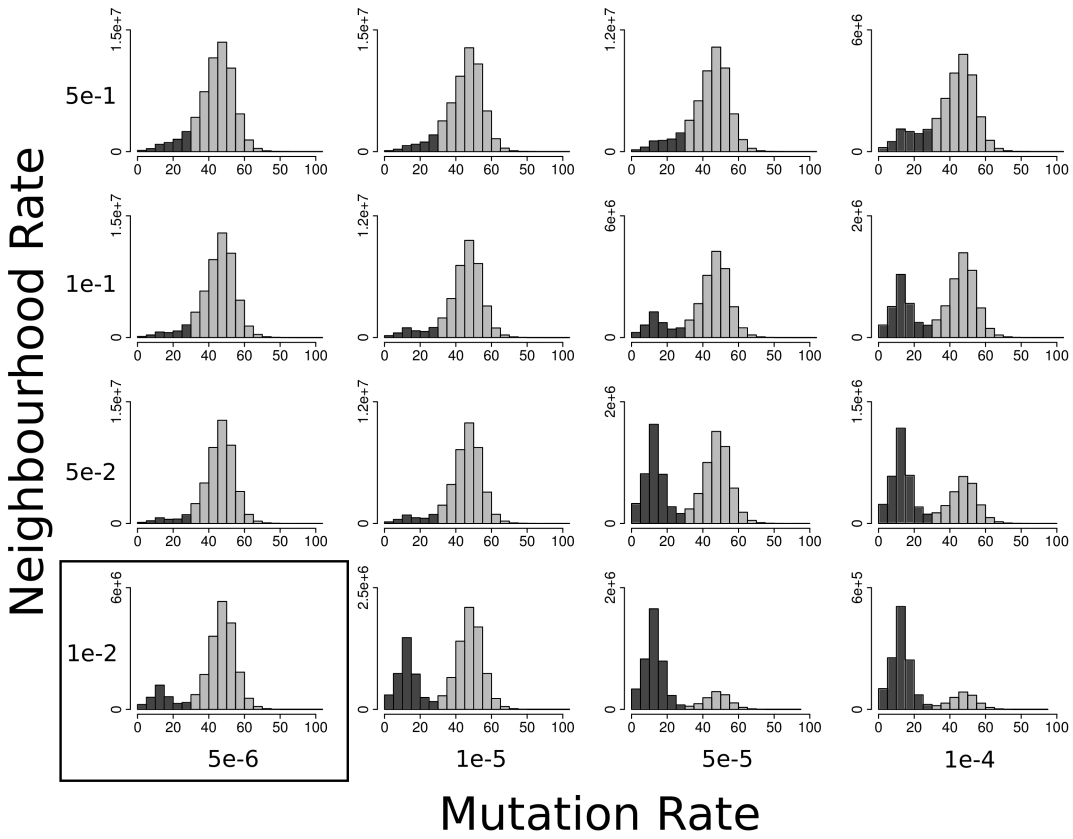


Figure IV.12 – Distribution of the scores of the alignments that caused a rearrangement to occur in the whole population and during the entire evolutionary process, for each value of μ_n and μ_m. Light grey: homologous rearrangements, dark grey: nonhomologous rearrangements. For computational performance reasons, the given values are minimal boundaries to the corresponding alignment score (cf. Algorithm 1).

very last generations, the events on this lineage are those that went to fixation, either by selection or by genetic drift. In addition, every other 10 generations, we used the standard bioinformatics tool Mummer (Kurtz et al., 2004) to find the most significant repeated sequences in the ancestral genomes. Mummer uses an approach similar to that of BLAST: it first searches for exact short repeats and then tries to join them together, allowing for gaps and mismatches. An example of Mummer output is shown in Figure IV.13. In this example, there are both direct and inverted repeats, and most of the repeated sequences are located in non-coding parts of the genome. This suggests that non-coding DNA plays a major role in genome evolvability by providing breakpoints for chromosomal rearrangements. The emergence of repeated sequences having little or no direct impact on fitness has already been observed in genetic programming (Langdon and Banzhaf, 2008) although in that particular case, these repeated sequences could be thought to participate in robustness rather than evolvability.

Figure IV.14 shows the results of the analysis of the whole lineage of ancestors. It shows that fitness improvements are strongly correlated with the presence of repeats in the genome and, consequently, with the occurrence of chromosomal rearrangements. The im-

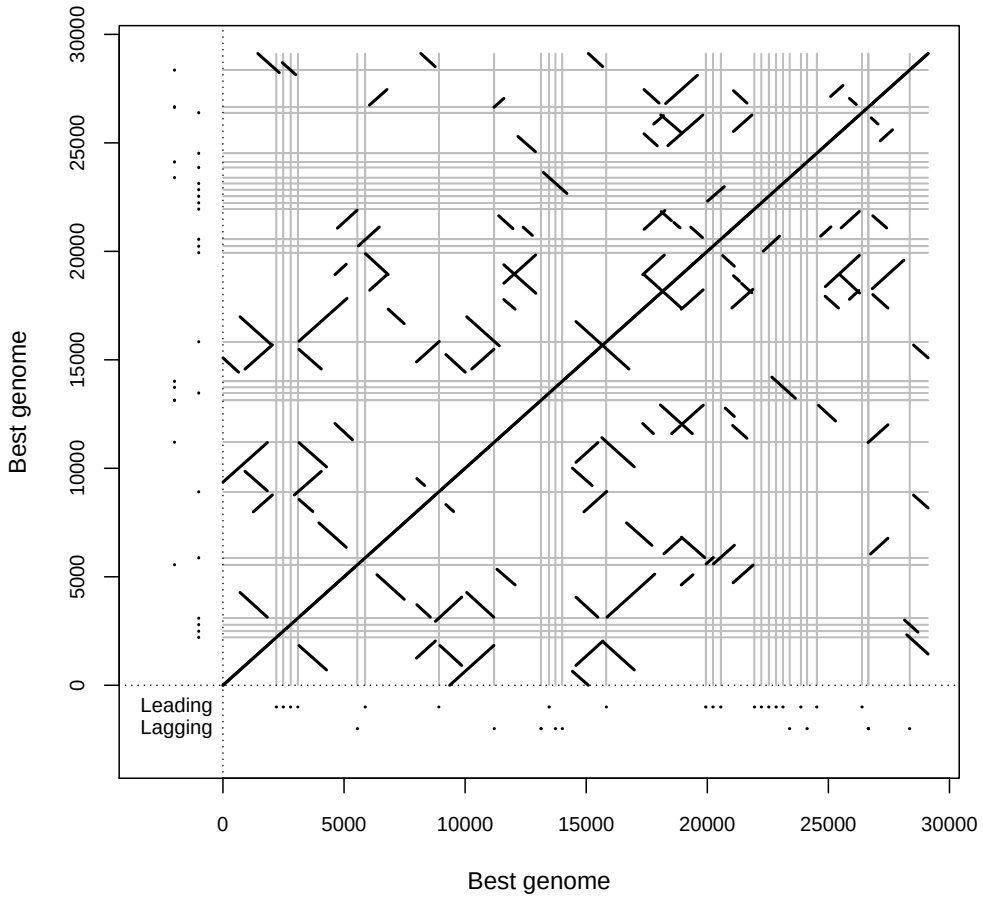


Figure IV.13 – Example of Mummer "dot plot" for the best individual at $t = 2000$ generations, for $\mu_n = 10^{-2}$ and $\mu_m = 5 \times 10^{-6}$, seed 2. Both the x- and the y-axis represent the genome of this individual. Long and strongly similar sequences appear as runs of diagonal lines across the matrix (exact match length = 15 bp, min. cluster length = 200 bp, max. gap between adjacent matches = 6 bp). Grey areas: coding sequences.

pact of chromosomal rearrangements on evolvability is thus rather complex: on the one hand, a very high rate of spontaneous rearrangements has a negative impact on the final fitness (Figure IV.11(c)), but on the other hand, in these simulations where the rate was low and the final fitness high, we find that the presence of rearrangements is correlated with fitness improvement (Figure IV.14). This suggests that a minimal amount of chromosomal rearrangements is required for evolution to be efficient.

A closer look at the rearrangements that went to fixation in these simulations (see Figure IV.15) reveals that *(i)* most of the fixed rearrangements were based on homologous breakpoints (*score* > 30), *(ii)* most of the fixed translocations and inversions were neutral, *(iii)* most of the fixed deletions were beneficial and *(iv)* most of the fixed duplications were deleterious. This last result is surprising at first sight: one would expect fixed events

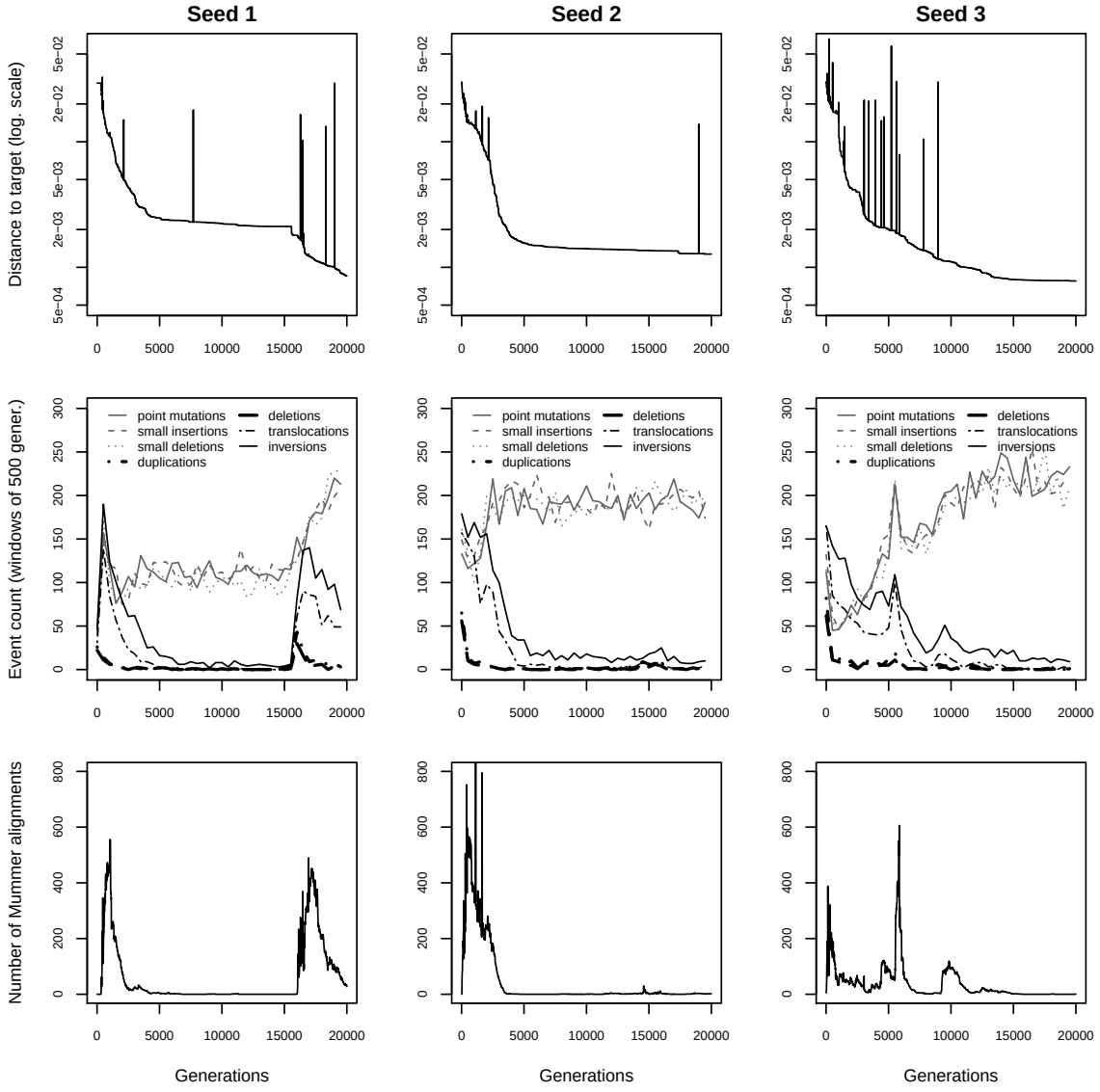


Figure IV.14 – Analysis of the line of ancestry of the final best individual for $\mu_n = 10^{-2}$ and $\mu_m = 5 \times 10^{-6}$. First row: evolution of fitness (the smaller the distance to the target, the higher the probability of reproduction). Second row: evolution of the number of mutational events, by windows of 500 generations. Third row: number of alignments found by Mummer on the genome (parameters: see Figure IV.13). Each column represents one of the three repetitions with the same set of parameters.

to be mostly neutral or beneficial. Our hypothesis is that despite their immediate negative impact, duplications can be indirectly selected because they allow for the creation of new gene copies (which can then undergo small mutations and ultimately realize new functions) and new repeats (which can then mediate other rearrangements). They can hence allow for subsequent fitness improvement, thus enabling their fixation. This effect is similar to the one observed by Adami (2006) where some deleterious mutations were observed in the line of descent of the eventual "winners". These deleterious mutations were relatively quickly followed by beneficial mutations and were often necessary for the

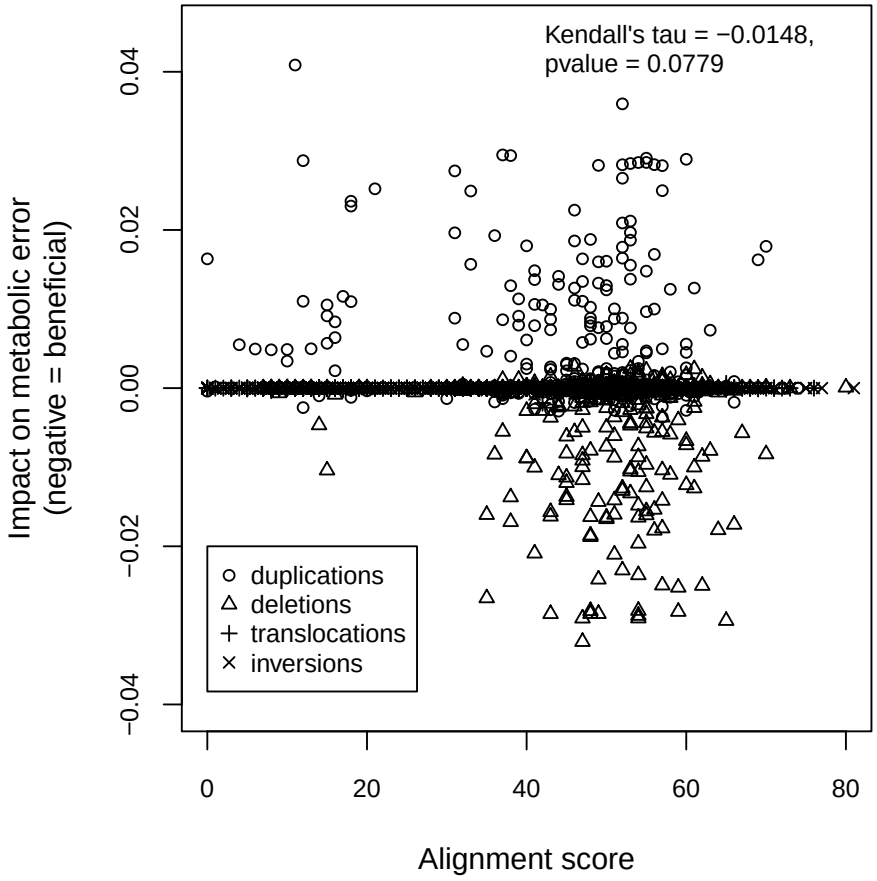


Figure IV.15 – Analysis of the fixed rearrangements for $\mu_n = 10^{-2}$ and $\mu_m = 5 \times 10^{-6}$ (all seeds together). Each point represents a rearrangement that occurred on the line of ancestry of the final best individual. The lower the point, the more beneficial the rearrangement. Interestingly, most beneficial rearrangements were homologous rearrangements ($score > 30$).

subsequent beneficial mutations to have an effect on fitness. A similar situation has been observed in *in vitro* experimental evolution where less fit but more evolvable organisms consistently prevailed in the long term (Woods et al., 2011).

6 Conclusion

In this chapter, we have proposed and validated a tractable model of homology-driven chromosomal rearrangements based upon an intermittent alignment search strategy. The results of our first set of experiments using this model confirm our previous results regarding the influence of rearrangements on genome compactness. In large genomes, repeated sequences (located mostly in non-coding regions) promote rearrangements that are, most of the time, deleterious. There is thus an indirect selective pressure to limit the number of rearrangements, which is done by eliminating repeats (fewer homologous rearrangements) and by reducing genome size (fewer nonhomologous rearrangements). However, we have

also shown that the absence of rearrangements is correlated with fitness stasis, suggesting that rearrangements can sometimes be directly beneficial or provide appropriate genetic background for subsequent beneficial mutations. A minimal amount of rearrangements is thus required for evolvability. Here, most of the rearrangements that went to fixation are homologous ones. For homologous rearrangements to be possible, repeats must be created at least as fast as they are destroyed by small mutations. In the end, the best conditions for evolvability seem to be a small basal rate of nonhomologous rearrangement combined with a low-enough mutation rate, thus leading to a few stable repeats and to an intermediate degree of variability by homologous rearrangements.

Chapter V

Horizontal Transfer

1 Introduction

Horizontal transfer (HT) is the transfer of genetic material between individuals that are not direct parents of one another. The terminology of *horizontal* transfer comes from the classical view of a tree of life where offspring are placed directly above or under their parents, individuals of the same generation thus sharing the same height on the tree. In this view, the transfer of genetic material between individuals of the same generation would appear as horizontal.

In bacteria, horizontal transfer can occur through different mechanisms, the most widespread being conjugation, transduction and transformation (Higgins, 2005; Lewin, 2007; Willey et al., 2007). Conjugation is an active process during which a bacteria injects a part of its genome into another bacteria by cell contact, usually through a "sex pilus". The most widely known example of conjugation is the F-factor transfer where a plasmid (secondary chromosome) called the F-factor is transferred from an F^+ bacteria (that owns a copy of the F-factor) to a former F^- bacteria (that doesn't own a copy of the F-factor), turning it into an F^+ bacteria (Willey et al., 2007). Transformation is the process by which a bacteria integrates some exogenous genetic material into its cell (*e.g.* DNA released in the environment by dead bacteria). Finally, transduction is a virus-mediated mechanism of horizontal transfer: a bacteriophage infecting a bacteria can make mistakes and pack some bacterial DNA into its capsid. Then, when the virus infects another bacteria, this DNA can be integrated to the genome of the newly infected bacteria, thus producing a horizontal transfer.

Horizontal transfer plays a major role in bacterial evolution, providing a way for bacteria to take advantage of beneficial mutations found by other bacteria, possibly from other species. Within a given species, horizontal transfers allow bacteria to evade the clonal interference phenomenon (Hill and Robertson, 1966) [1]: when two different beneficial mutations are found concomitantly in two different lineages, horizontal transfer allows both

1. The clonal interference phenomenon is also known as the Hill-Robertson effect.

mutations to be assembled into a single organism, thus speeding up evolution.
Another interesting possibility provided by horizontal transfer has been proposed, although indirectly, that regards evolvability through mutator alleles. It was shown that mutator alleles can accelerate adaptation to new environments by providing an alternative path towards adaptation through a transient mutator state: an antimutator can transiently turn into a mutator and turn back to an antimutator after having found a few beneficial mutations very quickly (Taddei et al., 1997; Tenaillon et al., 1999). Although in their model, a single mutational event with a fixed probability could revert a mutator to an antimutator, it might be thought that it would not be that simple for a mutator to re-discover *e.g.* an error-repair mechanism whose loss had led to its mutator state. However, horizontal transfer could well account for the reversion of these mutators to antimutators by reintegrating the lost allele from another antimutator lineage (provided that the mutator allele has not gone to fixation yet).
In such a model where evolvability is controlled by a specific locus, linkage disequilibrium is a necessary condition for second-order pressure to act upon this locus (hitchhiking). Then, if transfer is made available, the linkage disequilibrium can be broken, thus forbidding second-order selection. In Aevol however, the control of evolvability is distributed throughout the genome, so the second-order pressure on evolvability could persist even when transfer is allowed. We thus expect to observe the same kind of effects of the rearrangement rates on the size and structure of the genome as we have in previous experiments with no horizontal transfer.
To test this hypothesis, a biologically plausible – *i.e.* homology-driven – model of horizontal transfer was required. Transfers, like rearrangements, are potentially very dangerous, allowing for the replacement of any genetic sequence of one organism (the recipient) with any sequence (potentially of very different size and structure) from another organism (the donor). Now, biasing the process towards homologous rearrangements (as is the case in real organisms) should favour allelic recombination which can be thought as a lot less hazardous [1] while providing a way to evade linkage disequilibrium as well as clonal interference.

2 Mechanisms of Horizontal Transfer in bacteria

From the point of view of the sequence, conjugation, transformation and transduction all have their own specificities, both regarding the conditions necessary to their occurrence (whether it involves recombination and what kind of recombination) and their effects (sequence insertion or replacement).
Conjugation usually concerns plasmids, *i.e.* independent circular secondary chromosomes, so that no DNA recombination is required for such a transfer to take place, the plasmid being merely transferred from the donor to the recipient with no further interaction with other genetic material. Still, some plasmids can be integrated into the chromosome through a simple recombination event, involving a single pair of aligned sequences. It

1. As stated with humour by Nordin et al. (1999), "the natural exchange is strongly biased toward experimenting with features exchanging very similar chunks of the genome – specific genes performing specific functions – that have *small* variations among them, *e.g.* red eyes would be exchanged against green eyes, but not against a poor immune system".

may sometimes happen that part of the main chromosome gets transferred through conjugation, in which case the transferred segment is not circular. In such a situation, the transferred segment is single-stranded but is complemented upon arrival into the recipient cell (Willey et al., 2007). Then, recombination is needed for the segment to be integrated to the recipient's chromosome. In this specific case, since the segment is linear and is now double-stranded, two pairs of aligned sequences are needed for the transfer to succeed (reciprocal recombination).
In the case of transformation, the transferred sequence is single-stranded, and contrary to what happens during conjugation, it remains single-stranded in the recipient cell. Its integration into the recipient chromosome will then be achieved through non-reciprocal recombination, the single-stranded segment invading the chromosome and replacing one of the existing strands, thus forming heteroduplex DNA (Willey et al., 2007). This particular recombination requires the sequences to be similar along the whole length of the integrated segment.
Transduction is mediated by bacteriophage, a viral capsid conveying bacterial DNA due to a mishap during the virus life cycle. There are two types of transduction, generalized and specialized, both leading to similar outcomes. The phage containing DNA from the donor cell will inject it into the recipient cell as a double-stranded DNA fragment, possibly circular, that may subsequently be integrated into the recipient chromosome through reciprocal recombination (Willey et al., 2007).
To sum up, transduction leads to a transfer by replacement involving two breakpoints, one at each end of the transferred segment. Transformation also leads to a transfer by replacement but the transferred and replaced sequences must align along their whole length and not only at its extremities. Finally, conjugation can lead to either a transfer by insertion when a whole plasmid is transferred or by replacement when the transferred segment is not circular. Note that the latter case has both similar prerequisites and effects as transduction. Conjugation by insertion requires only one alignment for the former plasmid to be integrated into the main chromosome, and none at all if it remains a plasmid in the recipient cell.

3 Modelling Horizontal Transfer in Aevol

As we have seen in the previous section, three different mechanisms of horizontal transfer can lead to transfers either by replacement or by insertion and require either one or two local alignments or the alignment of the whole transferred sequence with the sequence to be replaced. Eventually, our goal is to study the dynamics of each kind of transfer. However, our first goal is to conduct a general study on the dynamics of transfer in general.
To introduce horizontal transfer in Aevol in a simple way, we modelled only one type of transfer. Since both transduction and conjugation can lead to a transfer by replacement requiring two local alignments (one at each end of the transferred segment) we assume this type of transfer to be most general and chose to model and study this particular kind of transfer first.
The transfer stage takes place at the beginning of replication. A new parameter μ_t will determine the proportion of replications during which a transfer will be attempted. For

those replications not falling in that proportion, the process is the same as in the former model. When a replication is "selected" for a transfer attempt, the recipient is selected using the exact same process as the parent selection process in the "classic" Aevol model (following the selection scheme – see chapter I section 6). A candidate donor (different from the recipient) is then selected at random in the population.
Then, as for the homology-driven rearrangement process presented in the previous chapter, a series of local searches will be performed between randomly chosen points from each of the genomes (donor and recipient), following a uniform distribution along the length of the corresponding genome. In other words, a random position is drawn in the genome of both the donor and the recipient and an alignment is searched for between the neighbouring sequences. The maximum number of candidate pairs of points to be tested is determined as the product of the *neighbourhood_rate* (parameter of the model – see chapter IV, section 4) and the size of the recipient's genome $nb\_pairs = \mu_n \times L_{recipient}$. The local search space is defined in the same way as for the homology-driven rearrangement process (see chapter IV, section 4).
A transfer will occur if two distinct alignments $A1$ and $A2$ are found between the chromosomes of the donor and of a recipient. If so, the entire segment defined between the breakpoints of A1 and A2 on the chromosome of the donor will replace the whole segment defined between the breakpoints of A1 and A2 in a copy of the chromosome of the recipient (*i.e.* in the chromosome of the offspring). If only one alignment was found (or none at all), the replication will go on without transfer.

4 Impact of Horizontal Transfer on Indirect Selection

Here, our goal is to explore the effects of horizontal transfer on evolution and in particular to test whether it alters the impact of chromosomal rearrangements on the structure of the genome that we have discussed throughout this manuscript. We let 140 populations of 1,000 individuals evolve during 50,000 generations in near identical conditions. The only changing parameters throughout the experiment were the rate μ_{mr} at which each type of both local mutations and chromosomal rearrangements occurred and the rate μ_t at which horizontal transfer events were attempted. We tested all the combinations of 7 different values for μ_{mr} (10^{-6}, 2×10^{-6}, 5×10^{-6}, 10^{-5}, 2×10^{-5}, 5×10^{-5} and 10^{-4} per bp per replication) and 4 different values for μ_t (0, *i.e.* no transfer at all, 10^{-3}, 10^{-2} and 10^{-1} per replication). Each combination was repeated five times with independent pseudorandom number generator seeds, yielding among other things a different initial population and different mutational events. The complete set of parameters used in these experiments is presented in table V.1. Note that in these experiments, the chromosomal rearrangements are realized regardless of the degree of similarity of the sequences (breakpoints randomly chosen as in the older version of Aevol). This is due to the excessive computational cost that would have resulted from the search for alignment both within the genome of each organism and between genomes in the same simulation. As was the case for the rearrangement rate in the experiments presented in the previous chapter, the rate at which transfer actually occurs is not fixed. It depends on both the parameter μ_t which determines the "transfer attempt" rate, and the presence of homologies between the genomes. However, as shown in figure V.1, the observed number of transfer events is strongly determined

by the rate at which they are attempted. Table V.2 shows, for each value of μ_{mr} in the simulations with $\mu_t = 10^{-1}$, the average proportion of replications in which a transfer actually occurred. We clearly see the effects of local mutations that make homologous sequences diverge in different lineages, thus reducing the actual rate of transfer. This is similar to what was observed for homologous rearrangements in the previous chapter.

Parameter	**Value**
N	1,000
nb_gener	50,000
$init\_length$	5,000
$init\_method$	Clonal, One Good Gene
$selection\_scheme$	Exponential Ranking
c	0.998
$E = \sum_i \alpha_i G_i$	$\alpha_1 = 1.2; G_1 : \mu = 0.52; \sigma^2 = 0.12$
	$\alpha_2 = -1.4; G_2 : \mu = 0.2; \sigma^2 = 0.07$
	$\alpha_3 = 0.3; G_3 : \mu = 0.8; \sigma^2 = 0.03$
$env\_sampling$	300
μ_{point}	$\mu_{mr} \in \{10^{-6}, 2\times10^{-6}, 5\times10^{-6}, 10^{-5}, 2\times10^{-5}, 5\times10^{-5}, 10^{-4}\}$
μ_{s_ins}	
μ_{s_del}	
μ_{dupl}	
μ_{del}	
μ_{inv}	
μ_{trans}	
$transfer\_attempt\_rate$	$\mu_t \in \{10^{-3}, 10^{-2}, 10^{-1}\}$
$p_{rear}(score)$	$\dfrac{1}{1+\exp-\dfrac{score-\alpha}{\lambda}}$ with $\alpha = 50$, $\lambda = 4$
$working\_zone\_half\_size$	50
max_shift	20
max_indel_size	6
W_{max}	0.01

Table V.1 – Parameters used in all the experiments of this chapter. The common rate for each kind of mutation and rearrangement takes its value among those proposed. The rate at which a transfer is attempted take its value among those proposed, a rate of 10^{-1} means that during one replication out of ten, a putative donor will be chosen at random in the population, the genomes of both the recipient and the donor will be searched for alignments, a transfer actually occurring if two alignments of sufficient score are found between the genomes.

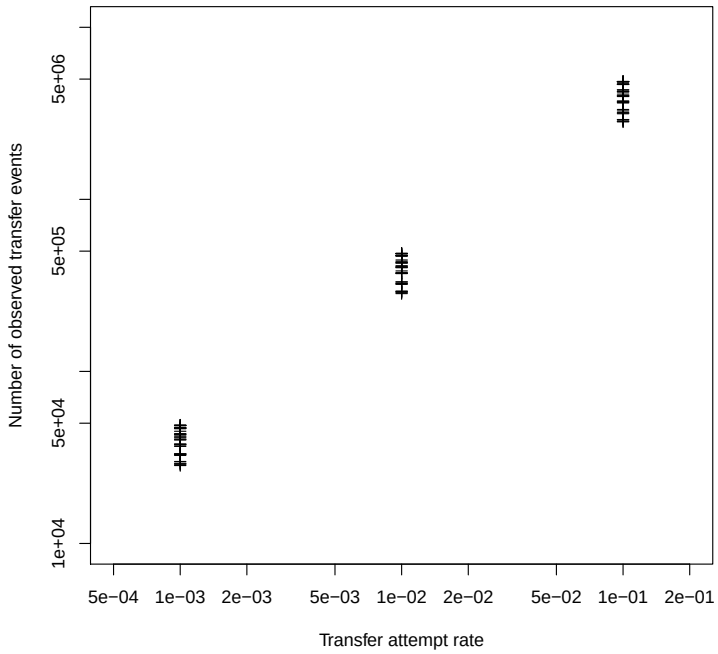
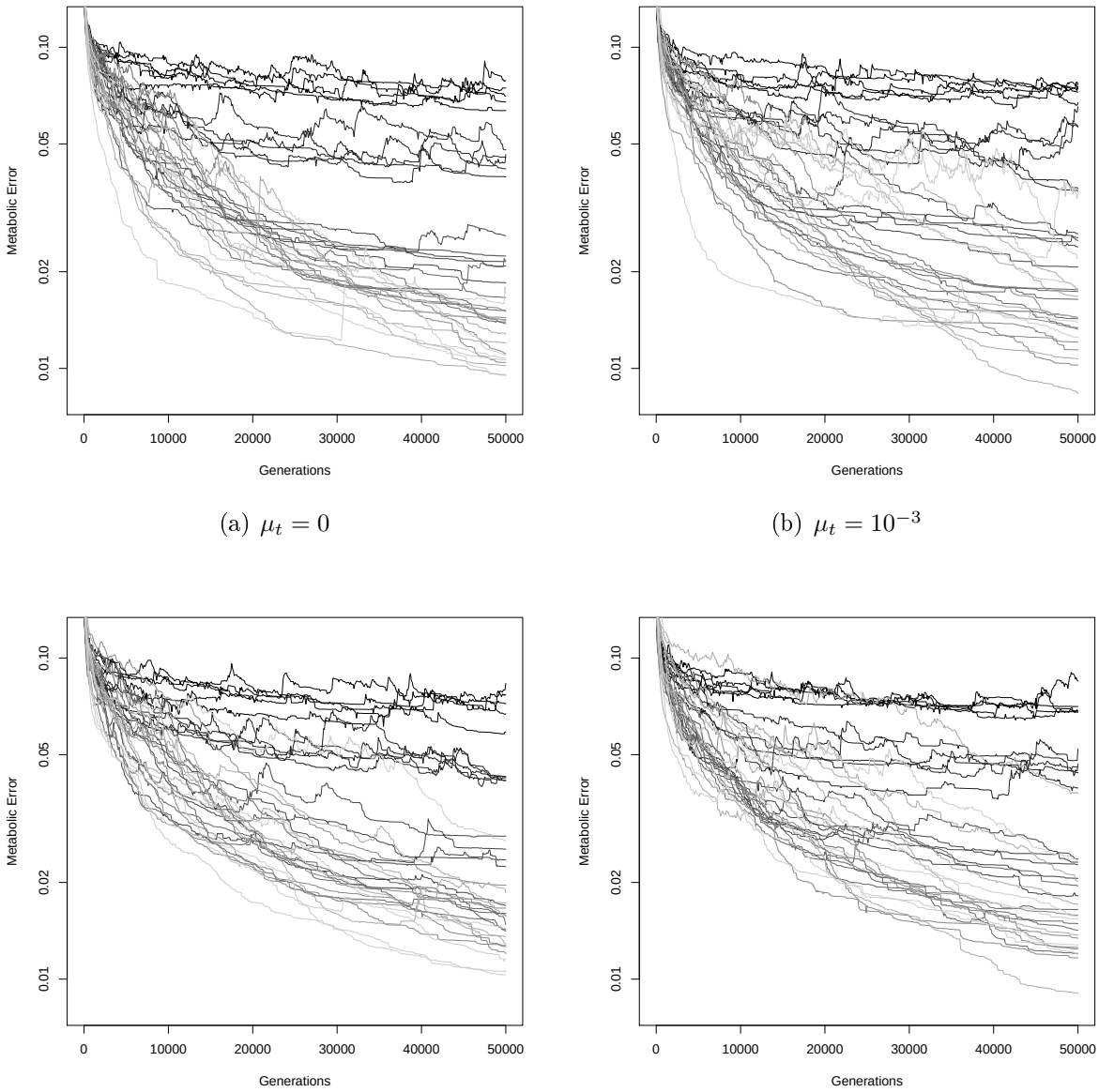


Figure V.1 – Number of transfer events observed during the whole evolution as a function of the transfer attempt rate μ_t.

10^{-4}	0.05729798
5×10^{-5}	0.06445151
2×10^{-5}	0.07364774
10^{-5}	0.08011026
5×10^{-6}	0.08534714
2×10^{-6}	0.09385118
10^{-6}	0.09670939

Table V.2 – Average proportion of replications involving a transfer event observed for each value of μ_{mr} in the simulations with $\mu_t = 10^{-1}$.

Figures V.2, V.3 and V.4 show respectively the evolution of the metabolic error, the number of genes and the number of non-coding bases of the best individual in the population throughout evolution for the different values of μ_t. Figure V.5 shows the last two indicators averaged over the last 1,000 generations. These figures suggest that horizontal transfer makes very little difference (if any) regarding the effects of the second-order pressure described in Chapter I, Section 9.2. This is very surprising since horizontal transfers, like chromosomal rearrangements, are large scale genetic variation events that could therefore well be thought to have dramatic effects on evolution.
The classical analysis of experiments conducted using Aevol involves the reconstruction of the lineage of the "winning" organism, thus providing us with important information

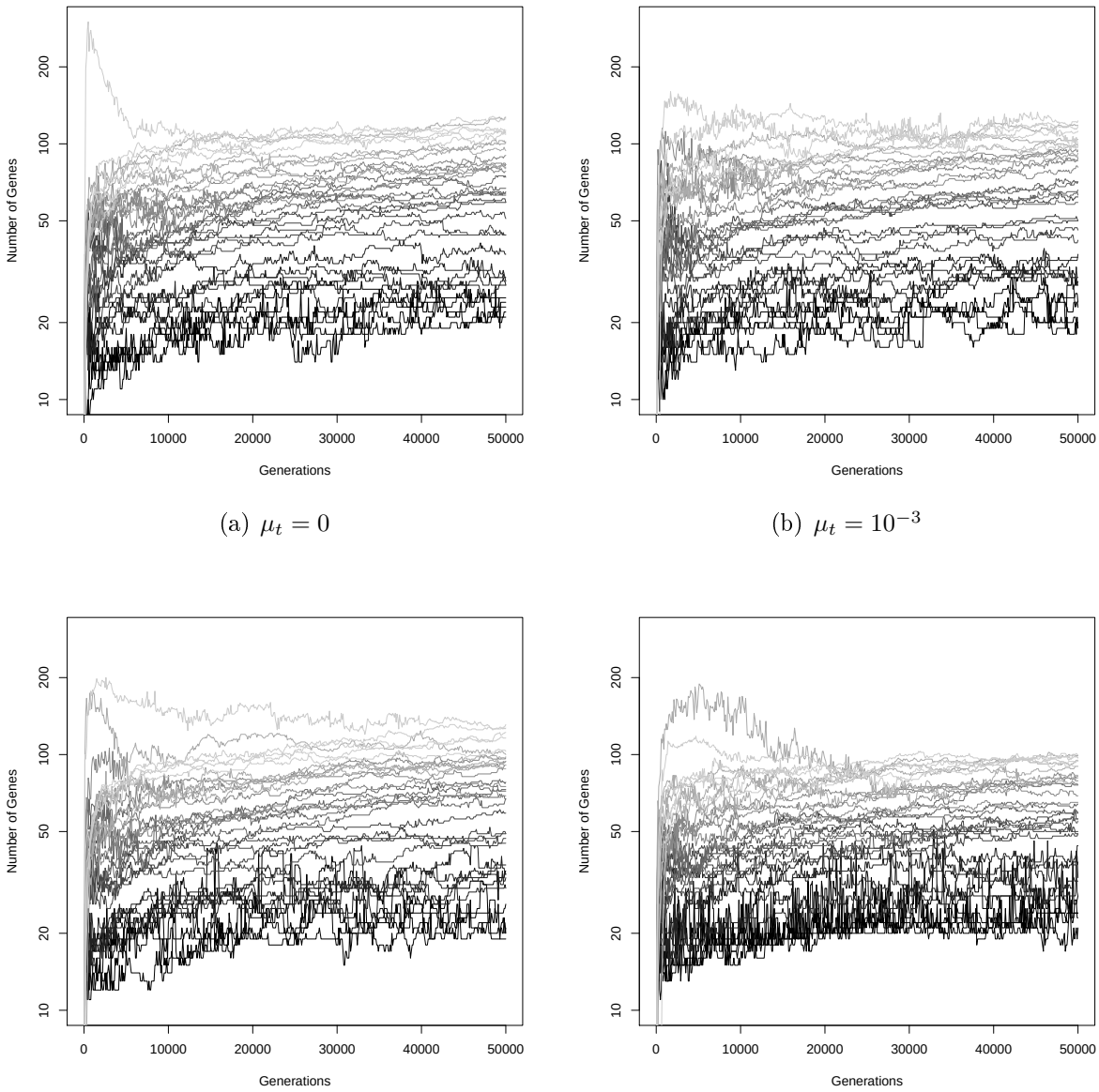


(a) $\mu_t = 0$ (b) $\mu_t = 10^{-3}$

(c) $\mu_t = 10^{-2}$ (d) $\mu_t = 10^{-1}$

Figure V.2 – Evolution of the metabolic error of the best organism of each generation. The grey scale corresponds to different values of μ_{mr}, black lines correspond to $\mu_{mr} = 10^{-4}$, the lightest grey to $\mu_{mr} = 10^{-6}$.

regarding the mutational events that went to fixation (see *e.g.* chapter IV, section 5). However, when horizontal transfer is allowed, the concept of the line of descent of an organism becomes at least unclear. Indeed, a horizontal transfer event can replace a segment of DNA of any size with another segment, itself of any size so that the genome of a given organism can be composed of half of the genomes of both of its ancestors. Then a "lineage" would be at best very "bushy" and very difficult to interpret. We hence need to



(a) $\mu_t = 0$

(b) $\mu_t = 10^{-3}$

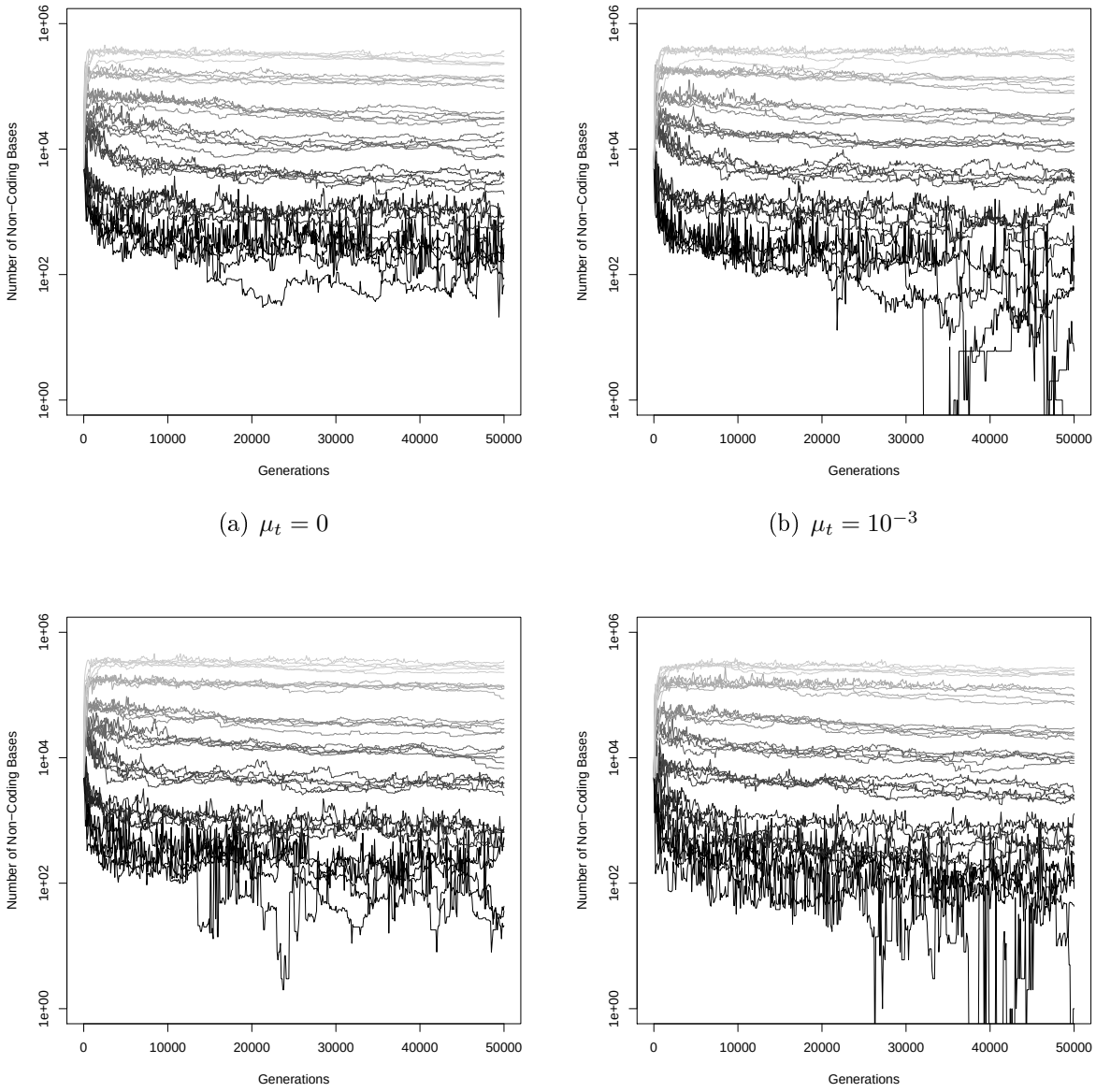
(c) $\mu_t = 10^{-2}$

(d) $\mu_t = 10^{-1}$

Figure V.3 – Evolution of the number of genes of the best organism of each generation. The grey scale corresponds to different values of μ_{mr}, black lines correspond to $\mu_{mr} = 10^{-4}$, the lightest grey to $\mu_{mr} = 10^{-6}$.

resort to indirect indicators to understand the dynamics that led to this apparent lack of effects of horizontal transfers.

To test whether homology makes a difference, we conducted further experiments in which no alignment was needed for a horizontal transfer event to occur and compared the different outcomes. Since the effects of horizontal transfer seem quite mild, we focused on the set of parameters where the number of HT events was greatest, *i.e.* on the simulations



(a) $\mu_t = 0$

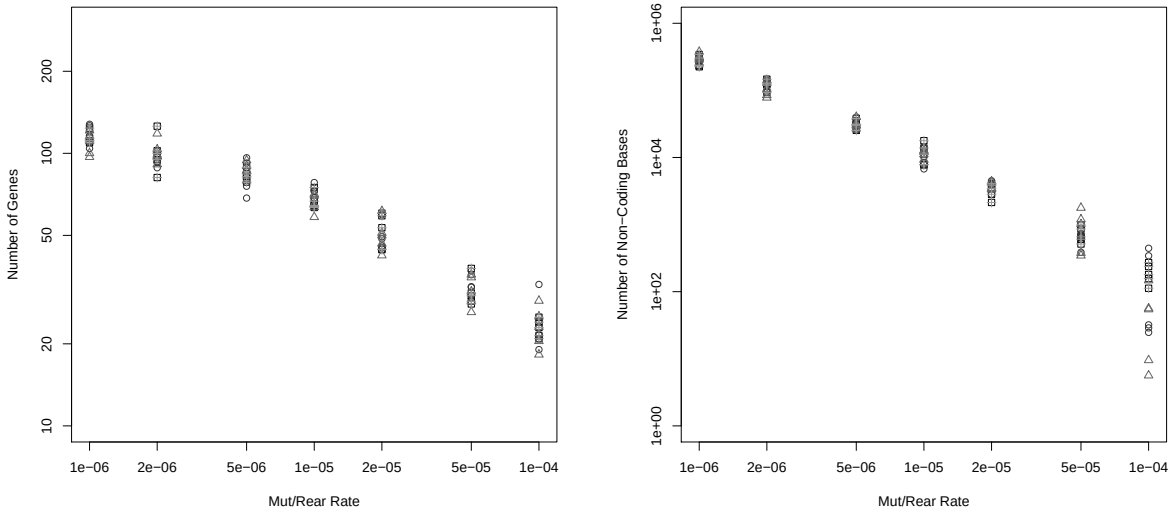
(b) $\mu_t = 10^{-3}$

(c) $\mu_t = 10^{-2}$

(d) $\mu_t = 10^{-1}$

Figure V.4 – Evolution of the amount of non-coding sequences of the best organism of each generation. The grey scale corresponds to different values of μ_{mr}, black lines correspond to $\mu_{mr} = 10^{-4}$, the lightest grey to $\mu_{mr} = 10^{-6}$.

where $\mu_t = 10^{-1}$. We simulated the evolution of 35 populations with a very similar set of parameters to that used for the main experiment: for each value of μ_{mr} tested, we injected into the new experiments the spontaneous rate of transfer that had been observed on average for the same value of μ_{mr} (see table V.2). This is because when no alignment whatsoever is needed for an HT to occur, every attempt leads to a transfer. In this particular case, μ_t *is* the spontaneous rate of transfer.



(a) Number of Genes

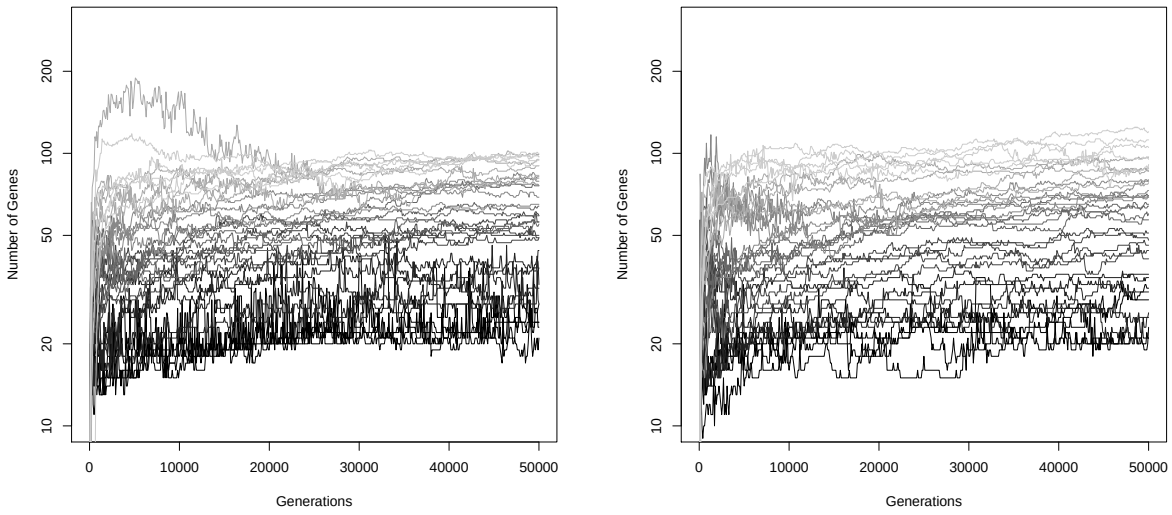
(b) Non-Coding Sequences

Figure V.5 – **(a)**: average number of genes and **(b)**: amount of non-coding sequences of the best organisms of the last 1,000 generations of each simulation.

Figures V.6, V.7 and V.8 show the number of genes and the number of non-coding bases of the best individual in the population throughout evolution and averaged over the last 1,000 generations for both the experiments where the HT process is sensitive to alignments and where it is random. It is again very surprising to observe that the sensitivity to sequence similarity in the choice for HT breakpoints seems to have very little impact (if any) on the structure of the evolved genomes.

However, a closer look to the horizontal transfer events show that there are indeed differences. Figure V.9 shows the proportion of replications involving an HT that was beneficial, deleterious or neutral. If there are no patent differences in the number of beneficial or deleterious events, it clearly appears that there are many more neutral events when homology matters. This could be the result of a greater probability for an exchange of homologous sequences: when sequence similarity (even circumscribed to the breakpoints) is needed for an exogenous sequence to replace part of the genome, it is more likely that the exogenote replaces its homologue than when the process is completely random. However, if this were the explanation for the observed difference in the number of neutral events, we would expect beneficial events to be more numerous in the case where HT occurs preferentially where sequences look alike, and this does not seem to be the case.

Beneficial mutations are rare. However, figure V.9 suggests that around one replication involving an HT out of ten is beneficial. This is due to the fact that, when HT is allowed, organisms with relatively low fitness can easily become better by replacing part of their genome by a homologous sequence taken from a fitter individual. If this is interesting *per se* on the level of the population, probably reducing its heterogeneity, it actually introduces a very strong bias when it comes to understanding the dynamics of novel mutations and their putative interferences. To reduce this bias, we focused on the replication of the



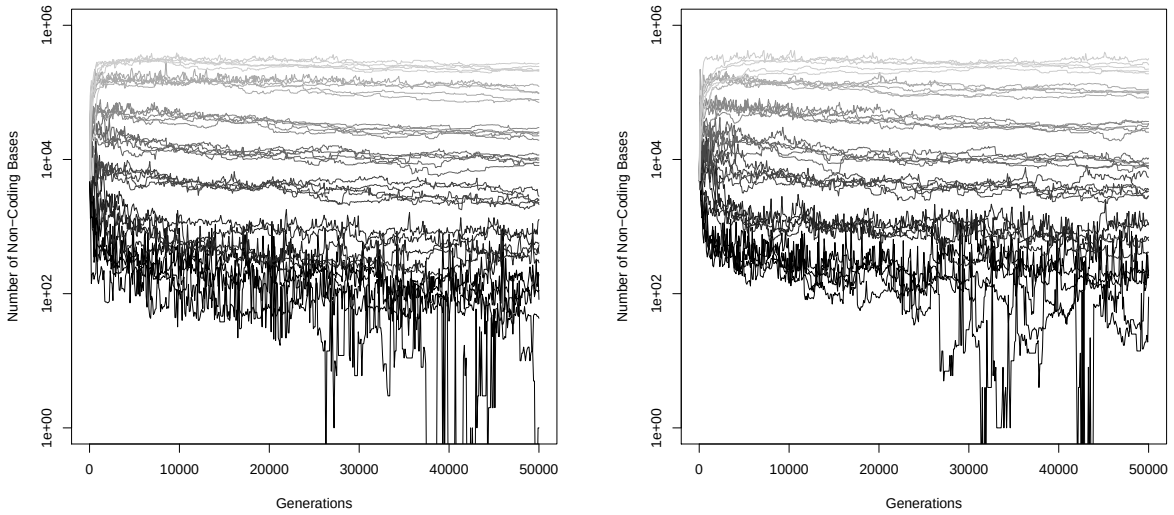
(a) With Alignments

(b) Without Alignments

Figure V.6 – Evolution of the number of genes of the best organism of each generation. **(a)**: simulation with homology driven transfer **(b)**: simulation with random point transfer. The grey scale corresponds to different values of μ_{mr}, black lines correspond to $\mu_{mr} = 10^{-4}$, the lightest grey to $\mu_{mr} = 10^{-6}$

top 100 individuals in the population. As shown in figure V.10, focusing on the replications where the recipient was already good before the transfer occurred reveals interesting differences in the proportion of neutral and beneficial events, both being more frequent when similarity is taken into account. This is particularly true at high mutation/rearrangement rates where homology-driven transfer proves to be both neutral and beneficial much more often than random transfer (over 10 times as much).

To better understand this effect, a closer look at the transfer events themselves is necessary. The size of the transferred and replaced sequences in particular are of great interest. Figure V.11 shows the average size of the transferred segments for beneficial, deleterious and neutral replications in both aligned and random transfer simulations. Note that since the genome is circular and the transferred segment between two breakpoints $bp1$ and $bp2$ can be either that going from $bp1$ to $bp2$ or from $bp2$ to $bp1$, the average size of the segments that are spontaneously transferred is roughly half the size of the donor's genome. It is worth noting that there are no clear differences between the experiments with or without alignments and that in both cases, transfers leading to neutral replications seem to involve smaller sequences. This is not very surprising. Indeed, large sequences are more likely to contain more genes and hence to cause more changes in the genome of the recipient than small sequences. Yet, the size of transferred segment is probably not the main determinant of the dangerousness of a transfer. Indeed, the transferred sequence can replace a segment of any size, potentially very different from its own size. So, even the transfer of a very small sequence could well have dramatic effects if it replaced a long sequence containing many genes. Following this idea, it is interesting to look at the relative sizes of the segment



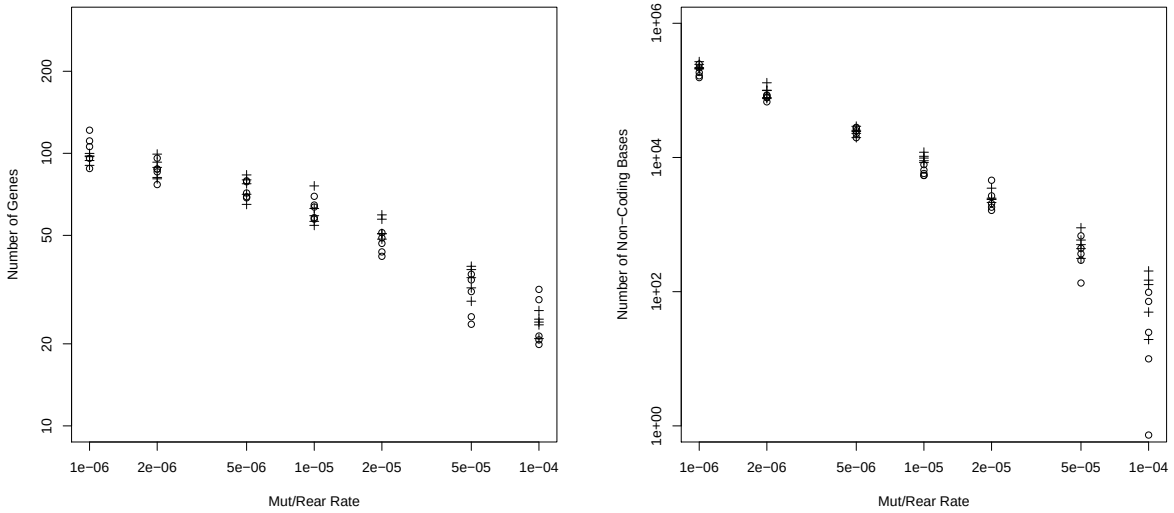
(a) With Alignments

(b) Without Alignments

Figure V.7 – Evolution of the amount of non-coding sequences of the best organism of each generation. **(a)**: simulation with homology driven transfer **(b)**: simulation with random point transfer. The grey scale corresponds to different values of μ_{mr}, black lines correspond to $\mu_{mr} = 10^{-4}$, the lightest grey to $\mu_{mr} = 10^{-6}$

that was transferred and the segment it replaced.

Looking at the relative sizes of the segment that was transferred and the replaced segments, the sensitivity to sequence similarity seems to favour those transfers whose segments are about the same size (figure V.12), again suggesting replacements between homologous sequences. This is even clearer when considering only very small differences in size: in figure V.13, we can observe that when alignments favour transfers, many transfers consist in replacing a given sequence by another sequence of exactly the same size. This reflects the replacement of a sequence by its homologue, possibly identical (neutral transfer) or having undergone base substitutions. We also observe that there are more transfers involving sequences that differ by only one to six bases in length than there are with greater differences. In this case, the transferred and the replaced sequences are homologous, one of them having undergone an indel (whose sizes are precisely comprised between one and six, see table V.1). Figure V.14 shows these differences in size between the transferred and the replaced segments in terms of proportion of the recipient's genome. This indicator will henceforth be referred to as δ_{size}. In this figure, each point corresponds to the average δ_{size} for beneficial, neutral or deleterious replications within a simulation. We can observe a great difference between both sets of simulations. Globally, δ_{size} seems to be smaller when alignments are accounted for. Beneficial events are of particular interest: at low mutation rates, the corresponding δ_{size} differs only slightly between the experiments with and without alignments. Above 2×10^{-5} however, the exchanged segments seem to be of roughly the same size with alignments while differing quite noticeably when alignments do not matter. This is all the more important since, as we have seen in figure V.10, there



(a) Number of Genes

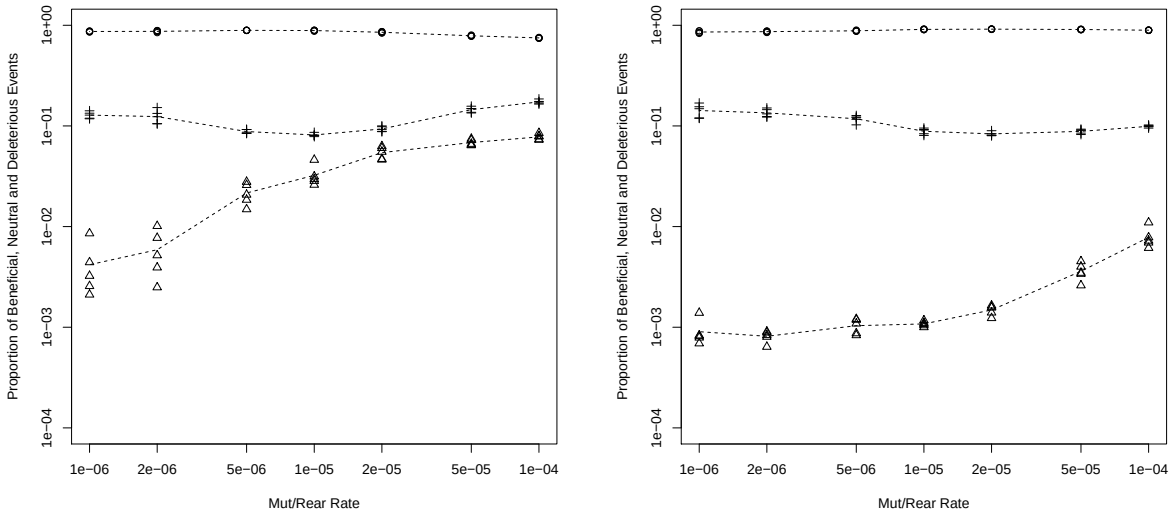
(b) Non-Coding Sequences

Figure V.8 – **(a)**: average number of genes and **(b)**: amount of non-coding sequences of the best organisms of the last 1,000 generations of each simulation.

are many more beneficial replications involving transfer at high mutation rates. The fact that neutral events correspond almost exclusively to sequences of the same size in the "aligned" case provides further confirmation that sensitivity to sequence similarity in the process of horizontal transfer favours allelic recombination.
This trend is confirmed by the analysis of the scores of the alignments that lead to a transfer in the simulations where homologies are accounted for. Figure V.15 shows that while both homologous and nonhomologous transfer can have deleterious effects, nonhomologous recombination is very unlikely to be neutral or to lead to a fitness improvement.

5 Discussion and Perspectives

In this chapter, we have presented a large scale experiment of digital evolution in which 175 populations of 1,000 individuals evolved independently for 50,000 generations. These 175 simulations fall into 5 different groups, each having a different scheme of horizontal transfer: groups A, B and C present different rates of transfer attempts, namely one attempt every 10, 100 or 1,000 replications respectively for groups A, B and C, a transfer actually occurring if two similar pairs of sequences are found between the genomes of the candidate donor and recipient. In group D, transfers were deterministically triggered between random points at the same rate as that effectively observed in group A. Finally, in group E, transfer was completely disabled. In each group, seven different values of mutation/rearrangement rate (μ_{mr}) were tested (the same rate for each type of mutations and rearrangements), namely 10^{-6}, 2×10^{-6}, 5×10^{-6}, 10^{-5}, 2×10^{-5}, 5×10^{-5} and 10^{-4} per base pair per replication.
The evolved organisms within each of these groups present very different structures, re-

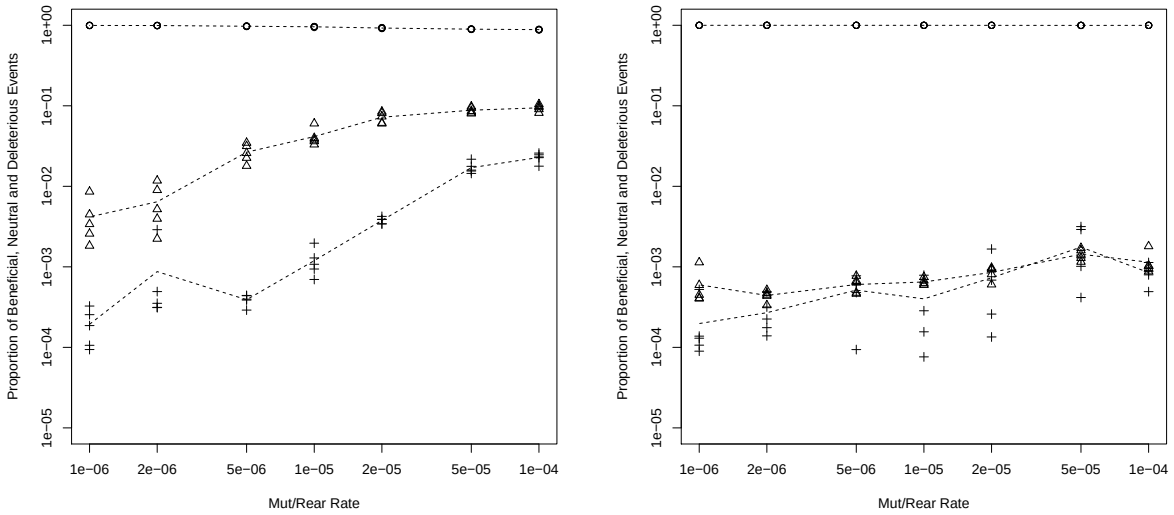


(a) With Alignments

(b) Without Alignments

Figure V.9 – Proportion of replications involving transfer that are beneficial (crosses), deleterious (circles) and neutral (triangles). **(a)**: simulation with homology driven transfer **(b)**: simulation with random point transfer.

producing very precisely the results we had previously obtained using the Aevol model without any kind of transfer: low values of μ_{mr} consistently lead to very large genomes containing many genes and a huge proportion of non-coding sequences while at high rates of mutations and rearrangements, the evolved genomes are very small, containing very few genes that are very tightly packed on the genome. It was shown that these differences in genome structure are due to a long term indirect selective pressure towards a specific trade-off between robustness and evolvability. The robustness and evolvability of an organism are usually thought to be the consequence of the specific set of genes or alleles it possesses, in particular the presence or absence of *e.g.* mutator alleles, some advanced error-repair mechanisms or chaperones. However, as we have previously stated, transfer provides a way to break linkage disequilibrium. Thus, a mutator that quickly found several beneficial mutations could recombine with an antimutator and become an antimutator itself. Thus, the mutator allele could not benefit from hitchhiking in the long term. Following this idea, allowing for transfer should prevent the second-order pressure on evolvability from being involved in the long term and the effects of this pressure on the structure of the genome that are observed when transfer is not allowed should not be observed when it is allowed. This, however, fails to happen here: evolution consistently shapes the genomes in the long term in such a way that they present a level of mutational variability close to the apparent optimum, even when horizontal transfer seems to enable the exchange of alleles between different lineages. This confirms that the level of mutational variability of a lineage is at least partly governed by the very structure of the genome. Indeed such an intrinsic, global and entangled factor of robustness or evolvability cannot be switched on or off by transfer, so that transfer does not yield any independence

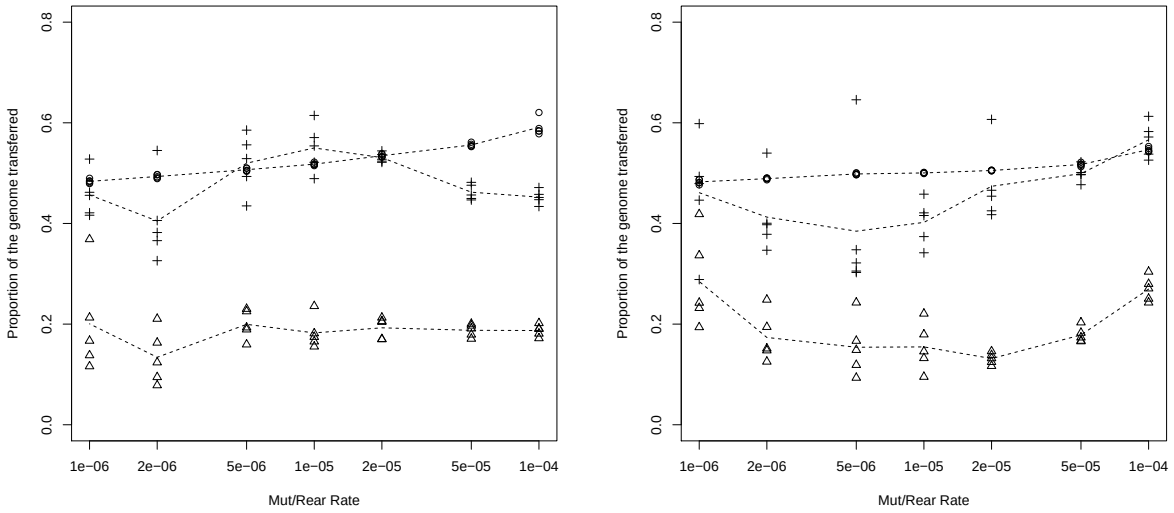


(a) With Alignments

(b) Without Alignments

Figure V.10 – Proportion of replications involving transfer among the best 100 organisms of each generation. Proportion of beneficial (crosses), deleterious (circles) and neutral (triangles) replications. **(a)**: simulation with homology driven transfer **(b)**: simulation with random point transfer.

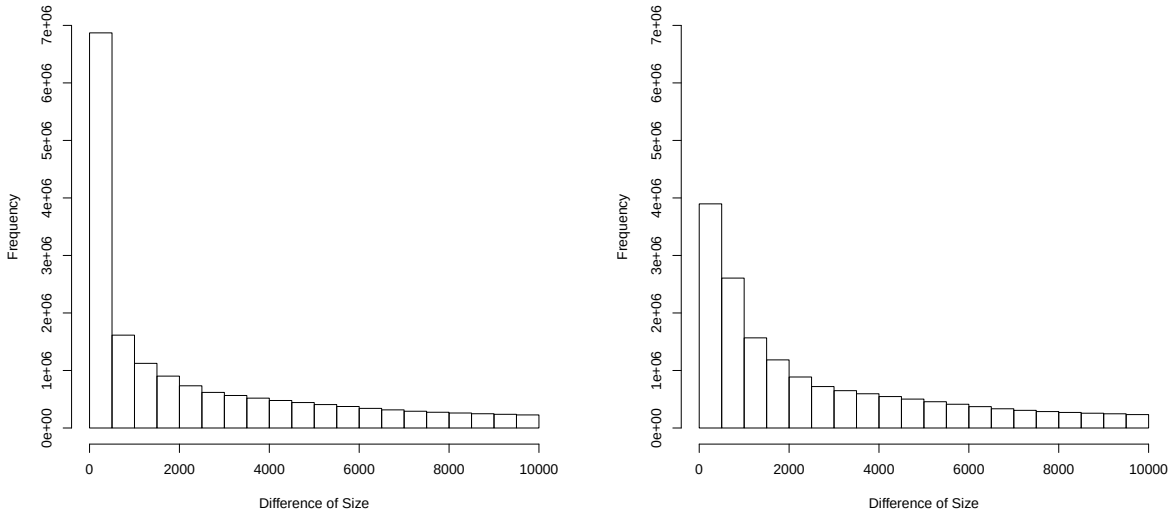
between these causes of evolvability or robustness and the level of variability itself.
The fact that no clear differences were identified regarding the fitness or the genome structure of the evolved organisms in these specific experiments does not mean that horizontal transfer has no advantages. As a matter of fact, the ubiquity of sex (in some form) in living systems strongly suggests that it is beneficial in some way. One of the possible explanations to the lack of differences in our experiments is the strong selective pressure that was applied to our organisms. The selection scheme that we used here (exponential ranking with $c = 0.998$) leads to short coalescence times (of the order of a few tens of generations) that might be too short for *e.g.* clonal interference to be a real problem. We hence look forward to testing horizontal transfer in different conditions, particularly under conditions of mild to moderate selection.



(a) With Alignments

(b) Without Alignments

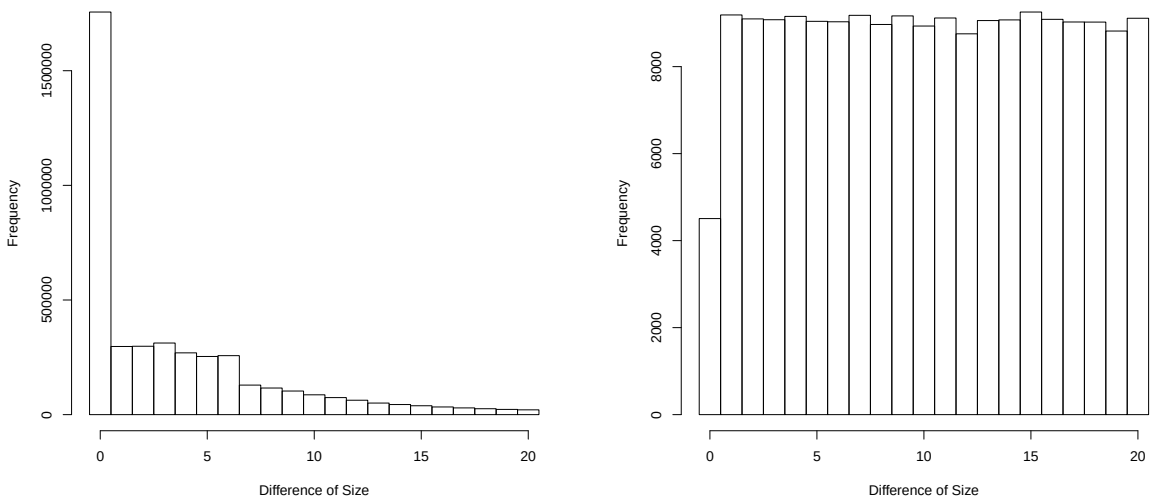
Figure V.11 – Proportion of the genome transferred during replications of the best 100 organisms of each generation for beneficial (crosses), deleterious (circles) and neutral (triangles) replications. **(a)**: simulation with homology driven transfer **(b)**: simulation with random point transfer.



(a) With Alignments

(b) Without Alignments

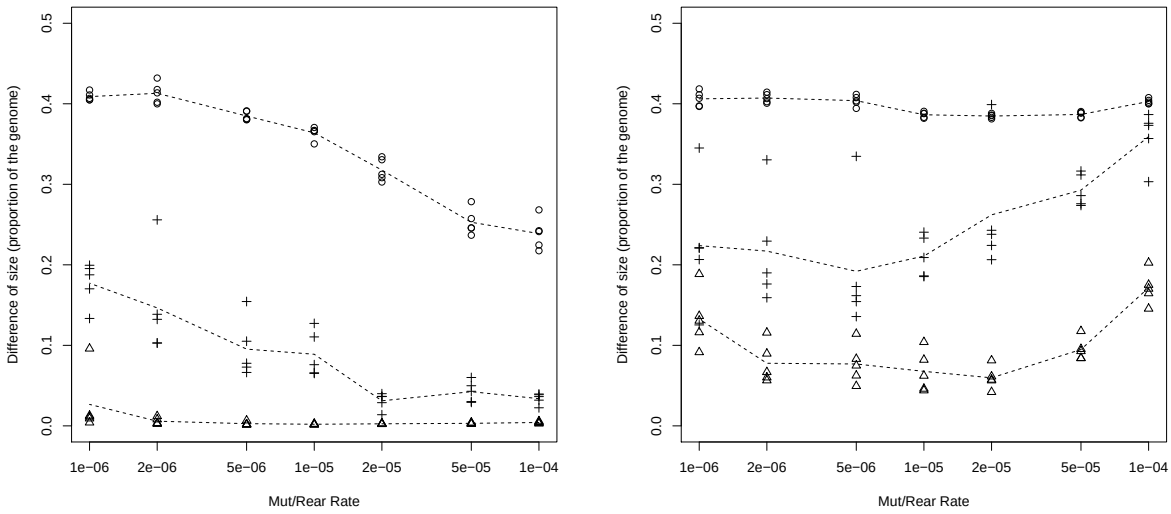
Figure V.12 – Distribution of the difference of size between the transferred and the replaced sequence during replications of the best 100 organisms of each generation for **(a)**: simulation with homology driven transfer and **(b)**: random point transfer.



(a) With Alignments

(b) Without Alignments

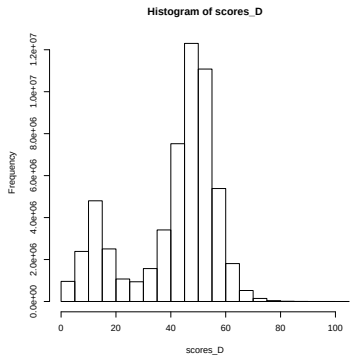
Figure V.13 – Detailed distribution for small difference of size between the transferred and the replaced sequence during replications of the best 100 organisms of each generation for **(a)**: simulation with homology driven transfer and **(b)**: random point transfer.



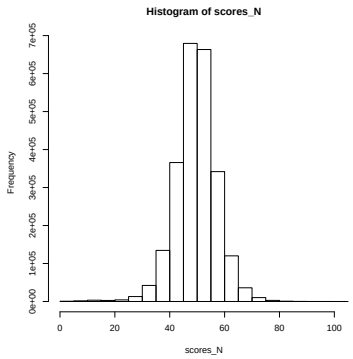
(a) With Alignments

(b) Without Alignments

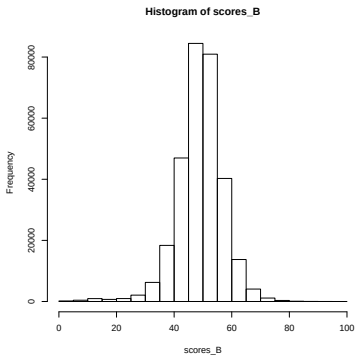
Figure V.14 – Difference of size (expressed in terms of proportion of the donor genome) between the transferred and the replaced sequence during replications of the best 100 organisms of each generation for beneficial (crosses), deleterious (circles) and neutral (triangles) replications. **(a)**: simulation with homology driven transfer **(b)**: simulation with random point transfer.



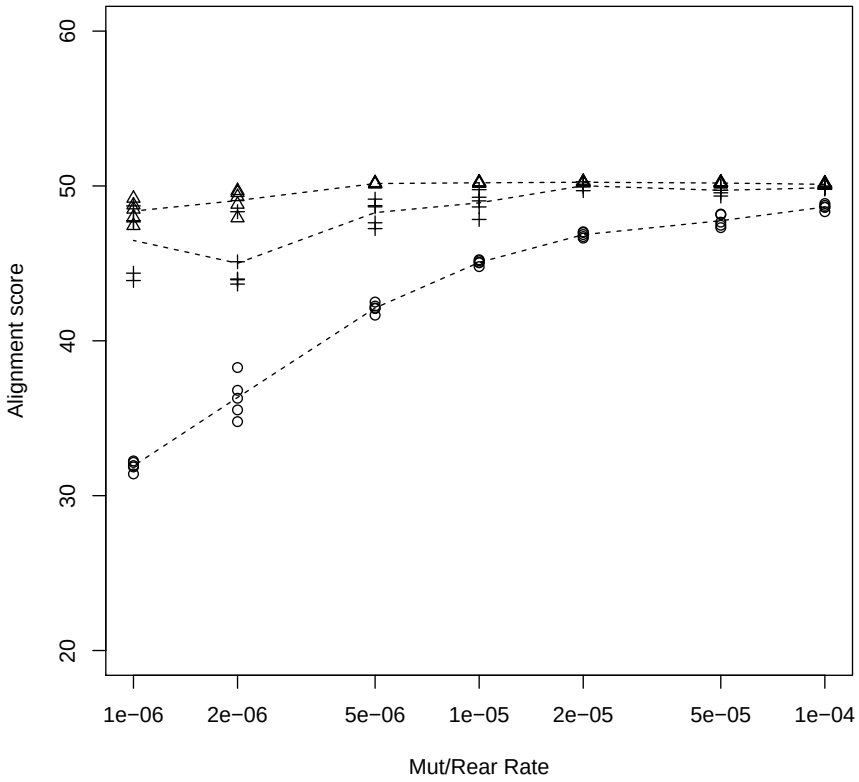
(a) Deleterious Events



(b) Neutral Events



(c) Beneficial Events



(d) Average alignment score between the breakpoints defining the transferred and replaced segments corresponding to deleterious, neutral and beneficial replications

Figure V.15 – Distribution of the score and average score of the alignments that lead to a transfer during replications of the best 100 organisms of each generation. **(a)**: deleterious, **(b)**: neutral and **(c)**: beneficial replication. **(d)**: average score for deleterious (circles), neutral (triangles) and beneficial (crosses) replications.

Conclusion

The Aevol model was designed by Carole Knibbe and Guillaume Beslon to study the evolution of genome structure. Using this model, they discovered a very strong indirect selective pressure toward a specific level of mutational variability of the phenotype. It was shown that, since non-coding sequences provide a substrate for chromosomal rearrangement breakpoints, these non-coding sequences are mutagenic for the genes they surround. Thus, when chromosomal rearrangements are involved, the per base-pair per replication rate at which they occur is a strong determinant of the structure of the genome: organisms having evolved in the context of low rearrangement rates usually own huge genomes with many genes and a lot of non-coding DNA while organisms having evolved with very high rates of rearrangement tend to have very short genomes with fewer genes and almost no non-coding DNA. Despite this great diversity of genome structure, all the evolved organisms have common characteristics: their genome is shaped by evolution in such a way that the best individual in the population tends to give birth to one offspring having exactly the same phenotype as its parent (*i.e.* one *neutral* offspring), suggesting the selection of a specific tradeoff between mutational robustness and evolvability in the long term.
These seminal results raised very interesting questions regarding both the specific conditions under which such a second-order pressure can be involved and the putative effects of this pressure on other levels of organization than the genome. To tackle these questions, the model was improved and extended to be both more efficient computationally and more realistic biologically. We also used the R-Aevol model, an extension of Aevol developed by Yolanda Sanchez-Dehesa during her PhD in which an explicit model of regulation of gene expression was introduced.
Given the central role of chromosomal rearrangements in the identified indirect pressure, a finer model of these rearrangements was needed to better understand the dynamics of this pressure. We hence improved the model to account for the specificities of chromosomal rearrangement mechanisms and in particular their sensitivity to sequence similarity. Finally, this model of homologous rearrangements was used to model horizontal transfer in a way that is biologically relevant.
We conducted experiments with both this extended model and with R-Aevol. In the experiment presented in chapter II, we tested different combinations of local mutation rates and of chromosomal rearrangement rates in the conditions of the former Aevol model (*i.e.* with no regulation nor transfer and where homologies were not accounted for) and studied the organisation of the transcriptome. The results show that while the local mutation rate has an impact mainly limited to the coding sequences, the rearrangement rate effectively applies a pressure on the entire genome, including non-coding sequences. Our results show that the second order pressure for evolvability that was identified by Knibbe et al. (2007a) has very interesting effects on the level of the transcriptome: these results showed that or-

ganisms having evolved under high rearrangement rates not only have very short genomes but also have very tightly packed genes which tend to be transcribed together on only a few polycistronic RNAs (operons), monocistronic RNAs and non-coding RNAs being seldom found on these genomes. On the contrary, organisms having evolved under low rearrangement rates and hence owning huge genomes usually have each of their genes transcribed on a different RNA (monocistronic). On these genomes, operons are rare and those few that can be found usually contain only a couple of genes. Non-coding RNAs however, are very numerous on these long genomes even though they have absolutely no function.

The experiments we conducted using the R-Aevol model showed that, despite the additional degree of freedom that is provided by regulation, the effects of the indirect selective pressure on the genome are very similar. Moreover, these effects come along with a clear trend for organisms having evolved in the context of low mutations and rearrangement rates to own very complex regulation networks, even when their environment is stable and hence regulation is not needed.

The experiments using the improved model, in which a rearrangement has a greater probability of occurring between similar sequences than between very different sequences, confirmed the results previously obtained regarding the effects of the spontaneous rate of rearrangements on the size and structure of the genome. The results also allowed us to identify a complex interaction between homologous rearrangements, nonhomologous rearrangements and local mutations and showed us that most beneficial rearrangements that had gone to fixation were homologous ones.

Finally, the last set of experiments we presented, in which homology-driven horizontal transfer was allowed, showed that even when transfer provides a way of evading the problem of linkage disequilibrium through allelic recombination, the effects of the studied pressure continue to strongly determine the structure of the genome. This confirms that, at least in the model, the structure of the genome itself has a great influence on the level of mutational variability of a lineage.

All these results (except for those regarding transfer) have been published independently. However, a global pattern persists throughout these experiments and it is always the same process that is involved: the second-order selection of a specific level of mutational variability of the phenotype. These results confirm those previously obtained by C. Knibbe and G. Beslon in a wide range of conditions, showing how robust these results are. Moreover, they show that the effects of this pressure span several levels of organization, including gene regulation networks. Overall, even though it is always fitness that drives evolution, it does so within the very strong constraints applied by the second-order selective pressure we have discussed.

This work shows that there is still much to be understood regarding the mechanisms that can lead to second-order selection and their effects. The different results we obtained in these relatively independent studies based on the Aevol model open great perspectives. In particular, it would be of great interest to conduct experiment crossovers to test for example the impact of homology-driven rearrangements or transfer in R-Aevol. These results point directly to two main directions for future work: the regulation of gene expression and horizontal transfer. Both directions will require the exploration of a wider range of parameters. The improvements of the model in terms of computational cost will allow us to conduct broader experiments using the R-Aevol model. This will enable us to test the

impact of the complexity of the environment on the evolved structures or the influence of regulation on the emergence of operons. Regarding transfer, it would be very interesting to explore under which conditions it is beneficial. We could then let organisms evolve the ability to transfer genetic material itself and explore in more detail the conditions under which "sex" might be selected for. Finally, transfer can itself give birth to other second-order pressures and is often mentioned as an explanation for the emergence of structures of interest (selfish operon, modularity). It would then be of great interest to study the effects of these pressures on the structure of the genome and the transcriptome.
Finally, as long as evolutionary biology is our main concern, an exciting perspective of this work is to compare the results obtained in the model with real organisms. It would be very interesting, for instance, to compare the genomes and transcriptomes that evolved in the model with the ever increasing amount of data available in public databases (in particular those regarding prokaryotes). Finally, *in silico* experimental evolution methods being very similar to those of *in vivo* or *in vitro* experimental evolution, comparing the dynamics observed either in the model or in real organisms would be fascinating and would enable us to progress in the validation of the Aevol and R-Aevol models, which has never been done in artificial evolution.

Bibliography

Adami, C. (2006). Digital genetics: unravelling the genetic basis of evolution. *Nat. Rev. Genet.*, 7(2):109–118.

Adami, C. and Brown, C. T. (1994). Evolutionary learning in the 2D artificial life systems avida. In Brooks, R. and Maes, P., editors, *Artificial Life IV*, pages 377–381. MIT Press.

Alon, U. (2003). Biological networks: the tinkerer as an engineer. *Science*, 301(5641):1866–1867.

Alon, U. (2007). Network motifs: theory and experimental approaches. *Nat Rev Genet*, 8(6):450–461.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402.

Azevedo, R. B. R., Lohaus, R., Srinivasan, S., Dang, K. K., and Burch, C. L. (2006). Sexual reproduction selects for robustness and negative epistasis in artificial gene networks. *Nature*, 440(7080):87–90.

Balleza, E., López-Bojorquez, L. N., Martínez-Antonio, A., Resendis-Antonio, O., Lozada-Chávez, I., Balderas-Martínez, Y. I., Encarnación, S., and Collado-Vides, J. (2009). Regulation by transcription factors in bacteria: beyond description. *FEMS Microbiology Reviews*, 33(1):133–151.

Banzhaf, W. (2003). Artificial regulatory networks and genetic programming. In Riolo, R. L. and Worzel, B., editors, *Genetic Programming Theory and Practice*, chapter 4, pages 43–62. Kluwer.

Barabási, A.-L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5(2):101–113.

Bénichou, O., Coppey, M., Moreau, M., Suet, P. H., and Voituriez, R. (2005). Optimal Search Strategies for Hidden Targets. *Physical Review Letters*, 94(19):198101+.

Bénichou, O., Loverdo, C., Moreau, M., and Voituriez, R. (2006). Two-dimensional intermittent search processes: An alternative to lévy flight strategies. *Phys. Rev. E*, 74(2):020102+.

Bénichou, O., Loverdo, C., Moreau, M., and Voituriez, R. (2011). Intermittent search strategies. *Review of Modern Physics*.

Beslon, G., Parsons, D., Pena, J. M., Rigotti, C., and Sanchez-Dehesa, Y. (2010a). From digital genetics to knowledge discovery: Perspectives in genetic network understanding. *Intelligent Data Analysis Journal*, 14(2):173–191.

Beslon, G., Parsons, D. P., Sanchez-Dehesa, Y., Pena, J., and Knibbe, C. (2010b). Scaling laws in bacterial genomes: A side-effect of selection of mutational robustness. *BioSystems*, 102(1):32–40.

Biebricher, C. and Eigen, M. (2005). The error threshold. *Virus Research*, 107(2):117–127.

Blickle, T. and Thiele, L. (1996). A comparison of selection schemes used in evolutionary algorithms. *Evol. Comput.*, 4(4):361–394.

Blount, Z. D., Borland, C. Z., and Lenski, R. E. (2008). Historical contingency and the evolution of a key innovation in an experimental population of escherichia coli. *Proceedings of the National Academy of Sciences*, 105(23):7899–7906.

Bongard, J. (2010). The utility of evolving simulated robot morphology increases with task complexity for object manipulation. *Artificial Life*, 16(3):201–223.

Bongard, J. and Paul, C. (2001). Making evolution an offer it can't refuse: Morphology and the extradimensional bypass advances in artificial life. In Kelemen, J. and Sosík, P., editors, *Advances in Artificial Life*, volume 2159 of *Lecture Notes in Computer Science*, chapter 43, pages 401–412. Springer Berlin / Heidelberg, Berlin, Heidelberg.

Bongard, J. C. and Pfeifer, R. (2003). *Evolving Complete Agents Using Artificial Ontogeny*, pages 237–258. Springer-Verlag.

Brown, Clifford, Liebovitch, Larry, Glendon, and Rachel (2007). Levy flights in dobe ju/hoansi foraging patterns. *Human Ecology*, 35(1):129–138.

Cases, I., de Lorenzo, V., and Ouzounis, C. A. (2003). Transcription regulation and environmental adaptation in bacteria. *Trends in Microbiology*, 11(6):248–253.

Chow, S. S., Wilke, C. O., Ofria, C., Lenski, R. E., and Adami, C. (2004). Adaptive radiation from resource competition in digital organisms. *Science (New York, N.Y.)*, 305(5680):84–86.

Ciliberti, S., Martin, O. C., and Wagner, A. (2007). Innovation and robustness in complex regulatory gene networks. *Proceedings of the National Academy of Sciences*, 104(34):13591–13596.

Coppey, M., Benichou, O., Voituriez, R., and Moreau, M. (2004). Kinetics of target site localization of a protein on DNA: A stochastic approach. *Biophys. J.*, 87(3):1640–1649.

Cordero, O. X. and Hogeweg, P. (2007). Large changes in regulome size herald the main prokaryotic lineages. *Trends Genet*, 23(10):488–493.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258):561–563.

Crombach, A. and Hogeweg, P. (2007). Chromosome rearrangements and the evolution of genome structuring and adaptability. *Molecular Biology and Evolution*, 24(5):1130–1139.

Crombach, A. and Hogeweg, P. (2008). Evolution of evolvability in gene regulatory networks. *PLoS Comp Biol*, 4(7).

Crombach, A. and Hogeweg, P. (2009). Evolution of resource cycling in ecosystems and individuals. *BMC Evolutionary Biology*, 9(1):122+.

Darwin, C. (1859). *On the Origin of Species by means of Natural Selection or the preservation of Favoured Races in the Struggle for Life*. Murray, John.

de Back, W. (2006). Eco-evolutionary experiments with situated agents. masters, Utrecht University.

Defoin Platel, M. (2004). *Homologie en Programmation GénétiqueApplication à la résolution d'un problème inverse*. PhD thesis, Université de Nice-Sophia Antipolis.

Defoin Platel, M., Clergue, M., and Collard, P. (2003). Maximum homologous crossover for linear genetic programming. In Ryan, C., Soule, T., Keijzer, M., Tsang, E., Poli, R., and Costa, E., editors, *Genetic Programming*, volume 2610 of *Lecture Notes in Computer Science*, chapter 18, pages 29–48. Springer Berlin / Heidelberg.

Devert, A., Bredeche, N., and Schoenauer, M. (2006). Blindbuilder: A new encoding to evolve Lego-Like structures genetic programming. In Collet, P., Tomassini, M., Ebner, M., Gustafson, S., and Ekárt, A., editors, *Genetic Programming*, volume 3905 of *Lecture Notes in Computer Science*, chapter 6, pages 61–72. Springer Berlin / Heidelberg, Berlin, Heidelberg.

Devert, A., Bredeche, N., and Schoenauer, M. (2007). Robust Multi-Cellular Developmental Design. In et al., D. T., editor, *Genetic and Evolutionary Computation COnference*, pages 982–989, London, Royaume-Uni. ACM SIGEVO, ACM Press.

D'haeseleer, P. (1994). Context preserving crossover in genetic programming. pages 256–261 vol.1.

Dittrich, P., Ziegler, J., and Banzhaf, W. (2001). Artificial chemistries-a review. *Artif Life*, 7(3):225–275.

Draghi, J. and Wagner, G. P. (2009). The evolutionary dynamics of evolvability in a gene network model. *Journal of Evolutionary Biology*, 22(3):599–611.

Drake, J. W. (1991). A constant rate of spontaneous mutation in dna-based microbes. *Proc Natl Acad Sci USA*, 88(16):7160–7164.

Drake, J. W., Charlesworth, B., Charlesworth, D., and Crow, J. F. (1998). Rates of spontaneous mutation. *Genetics*, 148(4):1667–1686.

Earl, D. J. and Deem, M. W. (2004). Evolvability is a selectable trait. *Proceedings of the National Academy of Sciences of the United States of America*, 101(32):11531–11536.

Edwards, A. M., Phillips, R. A., Watkins, N. W., Freeman, M. P., Murphy, E. J., Afanasyev, V., Buldyrev, S. V., da Luz, M. G. E., Raposo, E. P., Stanley, H. E., and Viswanathan, G. M. (2007). Revisiting lévy flight search patterns of wandering albatrosses, bumblebees and deer. *Nature*, 449(7165):1044–1048.

Eigen, M. (1971). Self-organization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, 58:456–523.

Eigen, M., Mccaskill, J., and Schuster, P. (1989). The molecular quasi-species. *Adv. Chem. Phys.*, 75:149–263.

Elena, S. and Sanjuan, R. (2008). The effect of genetic robustness on evolvability in digital organisms. *BMC Evolutionary Biology*, 8(1):284+.

Elena, S. F. and Lenski, R. E. (2003). Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nat. Rev. Genet.*, 4(6):457–469.

Espinosa-Soto, C. and Wagner, A. (2010). Specialization can drive the evolution of modularity. *PLoS Comput Biol*, 6(3):e1000719+.

Flamm, C., Ullrich, A., Ekker, H., Mann, M., Högerl, D., Rohrschneider, M., Sauer, S., Scheuermann, G., Klemm, K., Hofacker, I., and Stadler, P. (2010). Evolution of metabolic networks: a computational frame-work. *Journal of Systems Chemistry*, 1(1):1–14.

Foster, D. V., Kauffman, S. A., and Socolar, J. E. S. (2006). Network growth models and genetic regulatory networks. *Phys. Rev. E*, 73(3 Pt 1):031912.

François, P. and Hakim, V. (2004). Design of genetic networks with specified functions by evolution in silico. *Proc Natl Acad Sci USA*, 101(2):580–5.

Galli-Taliadoros, L. A., Sedgwick, J. D., Wood, S. A., and Korner, H. (1995). Gene knock-out technology: a methodological overview for the interested novice. *Journal of Immunological Methods*, 181(1):1–15.

Geier, F., Timmer, J., and Fleck, C. (2007). Reconstructing gene-regulatory networks from time series, knock-out data, and prior knowledge. *BMC systems biology*, 1:11.

Gu, Z., Steinmetz, L. M., Gu, X., Scharfe, C., Davis, R. W., and Li, W.-H. (2003). Role of duplicate genes in genetic robustness against null mutations. *Nature*, 421(6918):63–66.

Hanage, W. P., Spratt, B. G., Turner, K. M. E., and Fraser, C. (2006). Modelling bacterial speciation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361(1475):2039–2044.

Harding, S., Miller, J., and Banzhaf, W. (2010). Developments in cartesian genetic programming: self-modifying cgp. *Genetic Programming and Evolvable Machines*, 11:397–439. 10.1007/s10710-010-9114-1.

Hecker, M., Lambeck, S., Toepfer, S., van Someren, E., and Guthke, R. (2009). Gene regulatory network inference: data integration in dynamic models-a review. *BioSystems*, 96(1):86–103.

Hershberg, R., Yegerlotem, E., and Margalit, H. (2005). Chromosomal organization is shaped by the transcription regulatory network. *Trends Genet.*, 21(3):138–142.

Higgins, N. (2005). *The bacterial chromosome*. ASM Press.

Hill, W. G. and Robertson, A. (1966). The effect of linkage on limits to artificial selection. *Genetics Research*, 8(03):269–294.

Hindré, T., Knibbe, C., Beslon, G., and Schneider, D. (2011). New insights into bacterial life processes by experimental evolution strategies. *Nature Reviews Microbiology (in press)*.

Ideker, T., Winslow, L., and Lauffenburger, D. (2006). Bioengineering and systems biology. *Annals of Biomedical Engineering*, 34(2):257–264.

Itzkovitz, S., Tlusty, T., and Alon, U. (2006). Coding limits on the number of transcription factors. *BMC Genomics*, 7:239.

Jacob, F., Perrin, D., Sánchez, C., and Monod, J. (1960). L'opéron : groupe de gènes à expression coordonnée par un opérateur. *C. R. Acad. Sci. Paris 250*, pages 1727 – 1729.

Kapranov, P., Willingham, A. T., and Gingeras, T. R. (2007). Genome-wide transcription and the implications for genomic organization. *Nat. Rev. Genet.*, 8(6):413–423.

Kashtan, N. and Alon, U. (2005). Spontaneous evolution of modularity and network motifs. *Proc Natl Acad Sci USA*, 102(39):13773–13778.

Kashtan, N., Noor, E., and Alon, U. (2007). Varying environments can speed up evolution. *Proceedings of the National Academy of Sciences*, 104(34):13711–13716.

Kashtan, N., Parter, M., Dekel, E., Mayo, A. E., and Alon, U. (2009). Extinctions in heterogeneous environments and the evolution of modularity. *Evolution*, 63(8):1964–1975.

Kirschner, M. and Gerhart, J. (1998). Evolvability. *Proceedings of the National Academy of Sciences*, 95(15):8420–8427.

Knabe, J. F., Nehaniv, C. L., and Schilstra, M. J. (2008). Genetic regulatory network models of biological clocks: Evolutionary history matters. *Artificial Life*, 14(1):135–148.

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

Knibbe, C., Coulon, A., Mazet, O., Fayard, J.-M., and Beslon, G. (2007a). A long-term evolutionary pressure on the amount of noncoding DNA. *Mol. Biol. Evol.*, 24(10):2344–2353.

Knibbe, C., Fayard, J.-M., and Beslon, G. (2008). The topology of the protein network influences the dynamics of gene order: from systems biology to a systemic understanding of evolution. *Artificial Life*, 14(1):149–156.

Knibbe, C., Mazet, O., Chaudier, F., Fayard, J.-M., and Beslon, G. (2007b). Evolutionary coupling between the deleteriousness of gene mutations and the amount of non-coding sequences. *J. Theor. Biol.*, 244(4):621–630.

Komosiński, M. and Ulatowski, S. (1999). Framsticks: Towards a simulation of a Nature-Like world, creatures and evolution advances in artificial life. In Floreano, D., Nicoud, J.-D., and Mondada, F., editors, *Advances in Artificial Life*, volume 1674 of *Lecture Notes in Computer Science*, chapter 33, pages 261–265. Springer Berlin / Heidelberg.

Konstantinidis, K. T. and Tiedje, J. M. (2004). Trends between gene content and genome size in prokaryotic species with larger genomes. *Proc Natl Acad Sci USA*, 101(9):3160–3165.

Kuo, P. D., Banzhaf, W., and Leier, A. (2006). Network topology and the evolution of dynamics in an artificial genetic regulatory network model created by whole genome duplication and divergence. *BioSystems*, 85:177–200.

Kurtz, S., Phillippy, A., Delcher, A., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S. (2004). Versatile and open software for comparing large genomes. *Genome Biology*, 5:R12.

Langdon, W. and Banzhaf, W. (2008). Repeated patterns in genetic programming. *Natural Computing*, 7(4):589–613.

Langdon, W. B. (2000). Size fair and homologous tree crossovers for tree genetic programming. *Genetic Programming and Evolvable Machines*, 1(1):95–119.

Lawrence, J. (1999). Selfish operons: the evolutionary impact of gene clustering in prokaryotes and eukaryotes. *Curr. Opin. Genet. Dev.*, 9(6):642–648.

Lenski, R. E., Barrick, J. E., and Ofria, C. (2006). Balancing robustness and evolvability. *PLoS Biol*, 4(12):e428.

Lenski, R. E., Ofria, C., Pennock, R. T., and Adami, C. (2003). The evolutionary origin of complex features. *Nature*, 423(6936):139–144.

Lewin, B. (2007). *Genes IX*. Jones and Bartlett.

Lipman, D. J. and Pearson, W. R. (1985). Rapid and sensitive protein similarity searches. *Science*, 227(4693):1435–1441.

Luscombe, N. M., Qian, J., Zhang, Z., Johnson, T., and Gerstein, M. (2002). The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties. *Genome Biol*, 3(8):RESEARCH0040.

Lynch, M. (2006). Streamlining and simplification of microbial genome architecture. *Annu. Rev. Microbiol.*, 60(1):327–349.

Marbach, D., Schaffter, T., Mattiussi, C., and Floreano, D. (2009). Generating Realistic In Silico Gene Networks for Performance Assessment of Reverse Engineering Methods. *Journal of Computational Biology*, 16(2):229–239.

Martin, O. C. and Wagner, A. (2008). Multifunctionality and robustness Trade-Offs in model genetic circuits. *Biophysical Journal*, 94(8):2927–2937.

Maslov, S., Krishna, S., Pang, T., and Sneppen, K. (2009). Toolbox model of evolution of prokaryotic metabolic networks and their regulation. *Proc Natl Acad Sci USA*, page 6 p. (Epub ahead of print).

Mattick, J. S. and Makunin, I. V. (2006). Non-coding RNA. *Hum. Mol. Genet.*, 15 Spec No 1(suppl_1):R17–29.

Mattiussi, C. and Floreano, D. (2007). Analog Genetic Encoding for the Evolution of Circuits and Networks. *IEEE Transactions on Evolutionary Computation*, 11(5):596–607.

Maynard Smith, J. (1992). Byte-sized evolution. *Nature*, 355(6363):772–773.

Mendes, P., Sha, W., and Ye, K. (2003). Artificial gene networks for objective comparison of analysis algorithms. *Bioinformatics*, 19(suppl 2):ii122–ii129.

Misevic, D., Ofria, C., and Lenski, R. E. (2006). Sexual reproduction reshapes the genetic architecture of digital organisms. *Proc. R. Soc. B.*, 273(1585):457–464.

Molina, N. and van Nimwegen, E. (2008). The evolution of domain-content in bacterial genomes. *Biol Direct*, 3:51.

Molina, N. and van Nimwegen, E. (2009). Scaling laws in functional genome content across prokaryotic clades and lifestyles. *Trends Genet*, page 5 p. (Epub ahead of print).

Musso, F. and Feverati, G. (2011). Mutation-selection dynamics and error threshold in an evolutionary model for turing machines.

Needleman, S. and Wunsch, C. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453.

Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys Rev E*, 69(2):026113.

Nordin, P., Banzhaf, W., and Francone, F. (1999). Efficient evolution of machine code for cisc architectures using instruction blocks and homologous crossover. In *Advances in Genetic Programming 3, chapter 12*, pages 275–299. MIT Press.

Ochoa, G. (2006). Error thresholds in genetic algorithms. *Evolutionary Computation*, 14(2):157–182.

O'Neill, B. (2003). Digital evolution. *PLoS Biol*, 1(1):e18+.

Orgel, L. E. (1963). The maintenance of the accuracy of protein synthesis and its relevance to ageing. *Proceedings of the National Academy of Sciences of the United States of America*, 49:517–521.

Pál, C. and Hurst, L. D. (2004). Evidence against the selfish operon theory. *Trends Genet.*, 20(6):232–234.

Parsons, D. P., Knibbe, C., and Beslon, G. (2010a). Aevol : un modèle individu-centré pour l'étude de la structuration des génomes. In *MajecSTIC*.

Parsons, D. P., Knibbe, C., and Beslon, G. (2010b). Importance of the rearrangement rates on the organization of transcription. In *Proceedings of Artificial Life XII*, pages 479–486.

Parsons, D. P., Knibbe, C., and Beslon, G. (2011). Homologous and nonhomologous rearrangements: Interactions and effects on evolvability. In *Proceedings of the Eleventh European Conference on the Synthesis and Simulation of Living Systems (ECAL 11)*, pages 622–629.

Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences*, 85(8):2444–2448.

Poli, R. and Langdon, W. B. (1997). Genetic programming with one-point crossover and point mutation. In *Soft Computing in Engineering Design and Manufacturing*, pages 180–189. Springer-Verlag.

Pollack, J. B. and Lipson, H. (2000). The GOLEM project: Evolving hardware bodies and brains. *Evolvable Hardware, NASA/DoD Conference on*, 0:37+.

Ponjavic, J., Ponting, C. P., and Lunter, G. (2007). Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res.*, 17(5):556–565.

Ray, T. S. (1991). An approach to the synthesis of life. *Artificial Life II, Santa Fe Institute Studies in the Sciences of Complexity, vol. XI.*

Ray, T. S. (1992). Evolution, ecology and optimization of digital organisms. *Santa Fe Institute working paper 92-08-04.*

Rensing, C. (2002). The role of selective pressure and selfish dna in horizontal gene transfer and soil microbial community adaptation. *Soil Biol. Biochem.*, 34(3):285–296.

Reymond, N., Calevro, F., Viñuelas, J., Morin, N., Rahbé, Y., Febvay, G., Laugier, C., Douglas, A., Fayard, J.-M., and Charles, H. (2006). Different levels of transcriptional regulation due to trophic constraints in the reduced genome of *Buchnera aphidicola* aps. *Appl Environ Microbiol*, 72(12):7760–7766.

Richter, P. H. and Eigen, M. (1974). Diffusion controlled reaction rates in spheroidal geometry. application to repressor–operator association and membrane bound enzymes. *Biophysical chemistry*, 2(3):255–263.

Riggs, A. D., Bourgeois, S., and Cohn, M. (1970). The lac repressor-operator interaction. 3. kinetic studies. *Journal of molecular biology*, 53(3):401–417.

Rohmer, L., Fong, C., Abmayr, S., Wasnick, M., Freeman, T. L., Radey, M., Guina, T., Svensson, K., Hayden, H., Jacobs, M., Gallagher, L., Manoil, C., Ernst, R., Drees, B., Buckley, D., Haugen, E., Bovee, D., Zhou, Y., Chang, J., Levy, R., Lim, R., Gillett, W., Guenthener, D., Kang, A., Shaffer, S., Taylor, G., Chen, J., Gallis, B., D'Argenio, D., Forsman, M., Olson, M., Goodlett, D., Kaul, R., Miller, S., and Brittnacher, M. (2007). Comparison of francisella tularensis genomes reveals evolutionary events associated with the emergence of human pathogenic strains. *Genome Biology*, 8(6).

Sanchez-Dehesa, Y. (2009). *RÆvol : un modèle de génétique digitale pour étudier l'évolution des réseaux de régulation génétiques*. PhD thesis, INSA-Lyon.

Shlesinger, M. F. (2006). Mathematical physics: Search research. *Nature*, 443(7109):281–282.

Siegal, M. L. and Bergman, A. (2002). Waddington's canalization revisited: Developmental stability and evolution. *Proceedings of the National Academy of Sciences*, 99(16):10528–10532.

Sims, K. (1994a). Evolving 3d morphology and behavior by competition. *Artif. Life*, 1:353–372.

Sims, K. (1994b). Evolving virtual creatures. In *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, SIGGRAPH '94, pages 15–22. ACM.

Smith, T. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197.

Sniegowski, P., Gerrish, P., Johnson, T., and Shaver, A. (2000). The evolution of mutation rates: Separating causes from consequences. *Bioessays*, 22:1057–1066.

Stover, C. K., Pham, X. Q., Erwin, A. L., Mizoguchi, S. D., Warrener, P., Hickey, M. J., Brinkman, F. S., Hufnagle, W. O., Kowalik, D. J., Lagrou, M., Garber, R. L., Goltry, L., Tolentino, E., Westbrock-Wadman, S., Yuan, Y., Brody, L. L., Coulter, S. N., Folger, K. R., Kas, A., Larbig, K., Lim, R., Smith, K., Spencer, D., Wong, G. K., Wu, Z., Paulsen, I. T., Reizer, J., Saier, M. H., Hancock, R. E., Lory, S., and Olson, M. V. (2000). Complete genome sequence of *Pseudomonas aeruginosa* pa01, an opportunistic pathogen. *Nature*, 406(6799):959–964.

Struhl, K. (1999). Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. *Cell*, 98(1):1–4.

Taddei, F., Radman, M., Maynard-Smith, J., Toupance, B., Gouyon, P. H., and Godelle, B. (1997). Role of mutator alleles in adaptive evolution. *Nature*, 387:700–702.

Tenaillon, O., Toupance, B., Le Nagard, H., Taddei, F., and Godelle, B. (1999). Mutators, population size, adaptive landscape and the adaptation of asexual populations of bacteria. *Genetics*, 152(2):485–493.

Tusscher, K. T. and Hogeweg, P. (2009). The role of genome and gene regulatory network canalization in the evolution of multi-trait polymorphisms and sympatric speciation. *BMC Evolutionary Biology*, 9(1):159+.

Ullrich, A., Rohrschneider, M., Scheuermann, G., Stadler, P. F., and Flamm, C. (2011). In silico evolution of early metabolism. *Artificial Life*, 17(2):87–108.

van Nimwegen, E. (2003). Scaling laws in the functional content of genomes. *Trends Genet*, 19(9):479–484.

Viñuelas, J., Calevro, F., Remond, D., Bernillon, J., Rahbé, Y., Febvay, G., Fayard, J.-M., and Charles, H. (2007). Conservation of the links between gene transcription and chromosomal organization in the highly reduced genome of *Buchnera aphidicola*. *BMC Genomics*, 8:143.

Viswanathan, G. M., Afanasyev, V., Buldyrev, S. V., Murphy, E. J., Prince, P. A., and Stanley, H. E. (1996). Levy flight search patterns of wandering albatrosses. *Nature*, 381(6581):413–415.

Viswanathan, G. M., Buldyrev, S. V., Havlin, S., da Luz, M. G. E., Raposo, E. P., and Stanley, H. E. (1999). Optimizing the success of random searches. *Nature*, 401(6756):911–914.

von Hippel, P. H. and Berg, O. G. (1989). Facilitated target location in biological systems. *Journal of Biological Chemistry*, 264(2):675–678.

Wagner, A. (1996). Does evolutionary plasticity evolve? *Evolution*, 50(3):1008–1023.

Wilbur, W. J. and Lipman, D. J. (1983). Rapid similarity searches of nucleic acid and protein data banks. *Proceedings of the National Academy of Sciences*, 80(3):726–730.

Wilke, C. O. and Adami, C. (2003). Evolution of mutational robustness. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 522(1-2):3–11.

Wilke, C. O., Wang, J. L., Ofria, C., Lenski, R. E., and Adami, C. (2001). Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature*, 412(6844):331–333.

Will, S., Reiche, K., Hofacker, I. L., Stadler, P. F., and Backofen, R. (2007). Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput. Biol.*, 3(4):e65+.

Willey, J., Sherwood, L., and Woolverton, C. (2007). *Prescott/Harley/Klein's Microbiology*. McGraw-Hill Science/Engineering/Math, 7 edition.

Wolf, D. (2003). Motifs, modules and games in bacteria. *Current Opinion in Microbiology*, 6(2):125–134.

Woods, R. J., Barrick, J. E., Cooper, T. F., Shrestha, U., Kauth, M. R., and Lenski, R. E. (2011). Second-Order selection for evolvability in a large escherichia coli population. *Science*, 331(6023):1433–1436.

Yedid, G. and Bell, G. (2001). Microevolution in an electronic microcosm. *The American naturalist*, 157(5):465–487.

Zheng, Z.-M. M. and Baker, C. C. (2006). Papillomavirus genome structure, expression, and post-transcriptional regulation. *Frontiers in bioscience*, 11:2286–2302.

Zhu, X., Gerstein, M., and Snyder, M. (2007). Getting connected: analysis and principles of biological networks. *Genes & Development*, 21(9):1010–1024.

Appendix A

Aevol : un modèle individu-centré pour l'étude de la structuration des génomes

Le texte de cette annexe a partiellement été publié dans (Parsons et al., 2010a).

1 Introduction

S'il occupe une position centrale dans la biologie moderne, le processus d'évolution des espèces reste difficile à étudier. En effet, les mécanismes qui en sont responsables agissent sur des échelles de temps très longues, rendant difficile l'expérimentation directe sur l'objet d'étude [1]. À cet obstacle logistique, viennent s'ajouter des difficultés d'analyse des résultats : comment associer de façon claire un caractère observé à une cause évolutive précise lorsque l'on ne maîtrise pas l'ensemble des paramètres pouvant entrer en jeu?
Les approches de génomique comparative permettent de s'affranchir de ces contraintes de temps, cependant, elles sont basées sur un cliché instantané des séquences (l'ADN des espèces actuelles) et doivent *inférer* leur passé évolutif.
La génétique digitale est un nouveau champ de recherche qui a émergé dans les années 90. Des structures de données représentant des systèmes biologiques abstraits (les « organismes digitaux ») dans une « chimie artificielle » (Dittrich et al., 2001) sont interprétées par des programmes dédiés qui leur permettent de se dupliquer, de subir des mutations, d'évoluer et de s'adapter à leur environnement. D'un point de vue algorithmique, la génétique digitale est très proche des algorithmes évolutionnaires, cependant, le but n'est plus de trouver une solution à un problème d'optimisation spécifique mais d'étudier le processus évolutif lui-même dans une perspective de vie artificielle.
Les approches de génétique digitale permettent de réaliser facilement et en un temps raisonnable des expériences d'évolution contrôlées et reproductibles, donnant accès à un

1. Même pour des espèces à reproduction rapide comme les bactéries, une expérience directe d'évolution prend des dizaines d'années (Blount et al., 2008).

enregistrement fossile complet de l'intégralité du processus évolutif. Il est alors possible de soulever des questions essentielles quant aux fondements génétiques de l'évolution de la complexité, de l'organisation des génomes ou de la robustesse et de l'évolvabilité (Adami, 2006).
Nous présentons ici le modèle d'évolution expérimentale *in silico* Aevol ainsi qu'un aperçu des résultats obtenus grâce à ce modèle.

2 Le modèle Aevol

Aevol est un modèle d'évolution artificielle développé au LIRIS pour permettre l'étude de la structuration des génomes dans le processus évolutif. Largement inspiré de la génétique bactérienne, sa principale particularité vis-à-vis des algorithmes génétiques classiques est qu'il met en œuvre explicitement des processus de transcription et de traduction du génome, faisant ainsi apparaitre un niveau intermédiaire (le protéome) entre le génotype et le phénotype. Cette couche intermédiaire autorise une grande flexibilité au niveau du génome, permettant à la structure du génome elle-même d'évoluer. Nous pouvons alors interroger le modèle sur les causes possibles des différences observées chez les organismes réels en termes de structuration et d'organisation du génome.

2.1 Aevol - Principes généraux

Aevol est un modèle individu-centré qui permet de simuler l'évolution d'une population de N individus. Chaque individu possède un génome dont la structure est inspirée des génomes bactériens. Il s'agit d'une séquence binaire circulaire double-brin contenant un nombre variable de gènes séparés par des séquences non-codantes (figure A.1). Au début d'une simulation, tous les individus possèdent le même génome aléatoire, puis, à chaque génération, le génome de chaque individu est décodé pour déterminer son phénotype qui permettra à son tour de calculer la valeur d'adaptation (ou fitness) de l'individu.

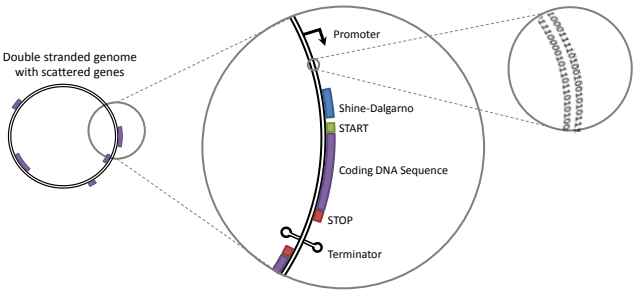


Figure A.1 – Dans Aevol, chaque organisme possède un génome binaire circulaire double-brin le long duquel des séquences codantes sont identifiées grâce à des séquences de signal prédéfinies. Les promoteurs et les terminateurs délimitent des séquences transcrites au sein desquelles des séquences codantes peuvent à leur tour être présentes entre un signal Start et un codon Stop en phase (voir figure A.2 pour le code génétique).

Plus un individu est adapté à son environnement, plus il a de chances de se reproduire. À

chaque génération, la population est intégralement renouvelée. Une roulette biaisée selon la fitness des individus permet de déterminer le « parent » de chaque nouvel individu.
Pendant le processus de réplication, le génome peut subir différents types de mutations : des mutations ponctuelles (substitutions, petites insertions ou délétions) mais aussi des réarrangements à l'échelle du chromosome (duplications, délétions, translocations, inversions). La structure du génome est donc libre d'évoluer (nombre de gènes, taille du génome, ...) et on peut étudier l'émergence de différentes structures génétiques.

2.2 Du génotype au phénotype

Dans Aevol, le décodage du génotype est directement inspiré des processus de transcription et de traduction bactériens. Nous avons défini un ensemble de signaux qui, lorsqu'ils sont présents sur l'ADN, nous permettent d'identifier les séquences qui seront transcrites en ARNs et, sur celles-ci, les sous-séquences qui seront traduites en protéines. Ces protéines seront ensuite interprétées en termes de « fonctions biologiques » réalisées ou inhibées par la protéine.

Transcription du génome

Chez les bactéries, l'initiation de la transcription s'effectue en des sites particuliers, appelés promoteurs, où les ARN-polymérases reconnaissent une séquence consensus et commencent la synthèse de l'ARN. Dans Aevol, un promoteur est une séquence dont la distance de Hamming d avec une séquence consensus prédéfinie, est inférieure ou égale à d_{max}. La séquence que nous utilisons typiquement dans nos expériences comporte 22 bases : 0101011001110010010110 et on autorise jusqu'à $d_{max} = 4$ différences. Cette séquence est suffisamment longue pour que des séquences non-codantes n'aient qu'une faible probabilité de devenir codantes à la suite d'une mutation.
Le niveau d'expression e d'un ARN dépend de sa séquence promotrice. Plus le promoteur est proche de la séquence consensus, plus le niveau d'expression est élevé : $e = 1 - \frac{d}{d_{max}+1}$. Cette modulation de l'expression des gènes modélise de façon simple l'interaction entre l'ARN-polymérase et le promoteur, sans introduire de réseau de régulation [1].
Lorsqu'un promoteur est identifié, la séquence est transcrite jusqu'à ce qu'un terminateur soit rencontré. Les terminateurs doivent être plus fréquents que les promoteurs pour limiter le chevauchement des séquences transcrites. Nous avons donc défini les terminateurs comme des séquences capables de former des structures en tige-boucle, similaires aux terminateurs ρ-indépendants bactériens [2]. Dans nos expériences, les tailles typiquement utilisées sont de 4 pour la tige et de 3 pour la boucle, ainsi les terminateurs ont la structure $abcd***\overline{d}\overline{c}\overline{b}\overline{a}$, où $a, b, c, d = 0$ ou 1.

Traduction des ARNs

Les séquences transcrites (ARNs) ne conduisent pas systématiquement à la production d'une protéine. Comme pour la transcription, le processus de traduction débute et se

1. Une extension du modèle (RAevol) intègre un mécanisme explicite de régulation de l'expression des gènes Beslon et al. (2010a,b).

2. Remarquablement, cette structure dite de « hairpin » permet de coder des terminateurs à la fois longs *et* fréquents.

termine lorsque le signal correspondant est rencontré. Ici, un signal de début de traduction est composé d'une séquence dite de Shine-Dalgarno, suivie, quelques bases plus loin, d'un codon Start (voir le code génétique figure A.2). Lorsque ce signal est rencontré, la séquence est lue codon par codon jusqu'à ce qu'un codon Stop soit trouvé dans le même cadre de lecture que le codon Start. Le processus de traduction associe alors à chaque codon (ou triplet de bases), un « acide aminé » abstrait grâce à un code génétique et la séquence d'acides aminés forme la séquence primaire de la protéine (figure A.2).

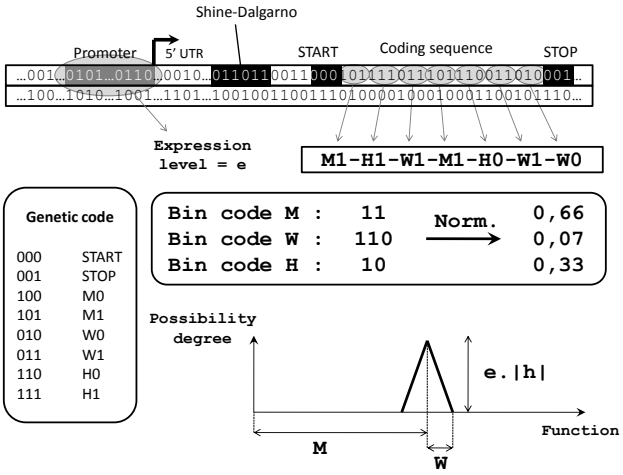


Figure A.2 – Schéma du processus de transcription-traduction-repliement dans Aevol. Les séquences transcrites sont celles qui commencent par un promoteur (séquence consensus) et finissent par un terminateur (structure tige-boucle), qui n'est pas sur la figure. Les séquences codantes (gènes) sont recherchées dans les séquences transcrites ; elles commencent par une séquence Shine-Dalgarno-START et se terminent par un codon STOP. Un code génétique artificiel (à gauche) est utilisé pour obtenir la séquence primaire de la protéine codée par un gène et un processus de « repliement » nous permet de calculer l'activité métabolique de la protéine (capacités fonctionnelles).

Comme chez les organismes réels, notre séquence génétique peut être lue suivant six cadres de lecture différents (trois sur chaque brin), ce qui permet aux organismes de présenter des gènes chevauchant (correspondant à des protéines différentes puisque lus sur des cadres de lecture différents).

Repliement des protéines et calcul du phénotype

Pour modéliser l'activité des protéines et le phénotype correspondant, nous avons défini une « chimie artificielle » simple (Dittrich et al., 2001) qui décrit le métabolisme d'un organisme dans un langage mathématique. Nous considérons qu'il existe un espace abstrait Ω de l'ensemble des processus métaboliques possibles. Dans le modèle, $\Omega = [0, 1]$, un processus métabolique est alors un simple réel. Dans cet « espace métabolique », chaque protéine est impliquée dans un ensemble de processus, soit en contribuant à leur réalisation, soit en les inhibant. Cette contribution est décrite grâce à un formalisme de logique floue : une protéine peut activer ou inhiber un processus biologique avec un degré de

possibilité compris entre 0 et 1 (positif pour une activation, négatif pour une inhibition). Une protéine est donc caractérisée par une fonction qui associe un degré de possibilité à chaque processus biologique. Pour des raisons de simplicité, nous utilisons des fonctions linéaires par parties ayant la forme de triangles isocèles (voir figure A.2). Ainsi, trois nombres suffisent pour caractériser l'activité métabolique d'une protéine : la position m ($m \in \Omega$) du triangle sur l'axe fonctionnel, sa demi-largeur w et sa hauteur h (positive quand la fonction est réalisée par la protéine, négative quand elle est inhibée). La protéine contribue donc à la plage de processus métaboliques $[m - w, m + w]$, avec une préférence pour les processus les plus proches de m (pour lequel la plus grande efficacité h est atteinte). Ainsi, plusieurs types de protéines peuvent co-exister, allant de protéines très spécialisées et efficaces (faible w, fort h) à des protéines beaucoup plus polyvalentes et moins efficaces (fort w, faible h).
Le calcul de ces trois paramètres à partir de la séquence primaire de la protéine est l'étape qui correspondrait dans les vraies cellules au repliement de le protéine. Ici la séquence primaire de chaque protéine est décomposée en trois sous-séquences binaires entrelacées codant les valeurs des trois paramètres m, w et h. Par exemple le codon 010 (resp. 011) est traduit en l'Acide Aminé $W0$ (resp. $W1$), ce qui signifie qu'il contribue au paramètre W en ajoutant un bit 0 (resp. 1) à son code binaire. La séquence binaire correspondant à chaque paramètre est finalement interprétée comme un réel normalisé selon la longueur de la séquence et les valeurs possibles du paramètre.
Une fois toutes les protéines d'un organisme identifiées et caractérisées, leurs activités respectives sont combinées en utilisant les opérateurs de Lucasiewicz. L'ensemble flou qui en résulte représente le phénotype P de l'individu, il indique le degré avec lequel cet individu réalise chaque fonction biologique de Ω.

2.3 Environnement, adaptation et sélection

Dans Aevol, l'environnement est représenté par une cible phénotypique : l'ensemble flou E défini sur Ω qui représente le degré de possibilité optimal pour chaque fonction biologique. Pour évaluer un individu, on compare son phénotype P à la cible E. L'aire géométrique g entre ces deux ensembles représente l'« erreur métabolique » de l'individu (figure A.3). Plus l'erreur métabolique est petite, meilleur est l'individu. Cette mesure pénalise aussi bien la sur- que la sous-réalisation de chaque fonction.
Chaque individu se voit attribuer une probabilité de reproduction en fonction de son erreur métabolique g et un tirage multinomial détermine le nombre de descendants effectif de chacun d'entre eux. Différentes méthodes de sélection sont disponibles, basés sur le rang de l'individu dans la population ou directement sur sa valeur d'adaptation (Blickle and Thiele, 1996). Toutes les expériences mentionnées ici ont été réalisées avec des sélections sur le rang, sans croisement entre les individus.

2.4 Opérateurs génétiques

Pendant leur réplication, les génomes peuvent subir sept types de mutations génétiques, parmi lesquels trois sont locaux (substitution d'une base, insertion ou délétion de quelques bases) et quatre sont des réarrangements chromosomiques affectant des segments potentiellement longs du génomes (duplication, délétion, translocation et inversion). Les points

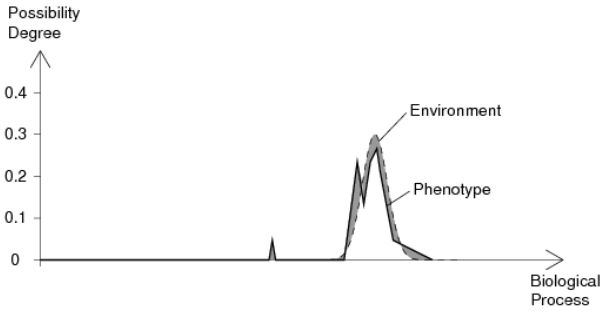


Figure A.3 – Mesure de l'adaptation d'un individu. Courbe pointillée : cible environnementale E. Courbe trait plein : phénotype P (profil métabolique obtenu en combinant toutes les protéines). Zone grisée : erreur métabolique g.

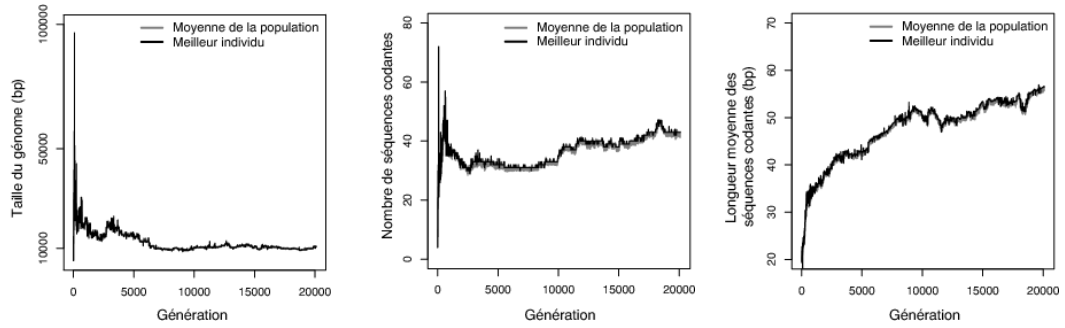
de rupture de ces réarrangements sont tirés aléatoirement sur le chromosome selon une loi uniforme.
Les mutations affectent le génome mais n'ont pas nécessairement un effet phénotypique. Ainsi, une mutation ayant lieu dans une région non transcrite sera complètement neutre (sauf si elle crée un nouveau gène, ce qui est très rare). Les taux de mutations μ_i sont des paramètres du modèle, ils sont définis comme la probabilité par base et par réplication qu'un évènement de type i ait lieu.

Aevol est donc un modèle de génétique digitale dans lequel la structure des génomes est libre d'évoluer. Il intègre les principaux mécanismes impliqués dans l'expression et la modification du génome, introduisant un niveau intermédiaire entre le génotype et le phénotype et autorisant non seulement des opérateurs de mutations ponctuelles, mais aussi les réarrangements chromosomiques.
Ces particularités font d'Aevol un modèle particulièrement adapté à l'étude de l'organisation des génomes. Il permet de réaliser des campagnes expérimentales complètes dans différentes conditions expérimentales (*e.g.* différents taux de mutations) et d'observer comment les paramètres structuraux des génomes évoluent en fonction de ces conditions. Il est alors possible de vérifier la cohérence des résultats obtenus avec les différentes hypothèses proposées dans la littérature et d'essayer de comprendre les mécanismes à l'origine des phénomènes observés.

3 Une évolution typique dans Aevol

Aevol permet de mener des campagnes d'évolution expérimentale sur plusieurs dizaines de milliers de générations et d'analyser l'allure des génomes obtenus en fonction des paramètres. Si les structures finales peuvent être très différentes, le processus évolutif est quant à lui relativement stable d'une expérience à l'autre.
On observe ainsi une amélioration rapide de la fitness des individus dans les premières générations puis un ralentissement progressif. On notera que la fitness n'est jamais totale-

ment stable et que des mutations avantageuses se produisent régulièrement, même après 500 000 générations.
L'évolution des individus s'accompagne de profondes modifications dans la structure de leur génome (figure A.4). Dans un premier temps, la taille du génome augmente fortement pour passer des 5 000 paires de bases initiales (initialisation par défaut dans aevol) à plusieurs dizaines de milliers de paires de bases en quelques centaines de générations. Le nombre de gènes et la taille des séquences non codantes augmentent aussi fortement. La deuxième phase se caractérise par une décroissance rapide de la taille du génome et du nombre de gènes, tandis que la taille des gènes, elle, continue de croître. Enfin, au cours de la troisième phase, la taille des génomes est stable. Par contre, l'organisme recommence à acquérir des gènes (mais plus modérément) tandis que la taille des séquences codantes augmente continûment.
Ainsi, dans un premier temps, les organismes augmentent rapidement la taille de leur répertoire génique, le plus souvent par duplication-divergence de gènes pré-existants. Ils sélectionnent ensuite les gènes les plus adaptés avant d'affiner leur répertoire génique en améliorant progressivement chacune de leurs séquences codantes.

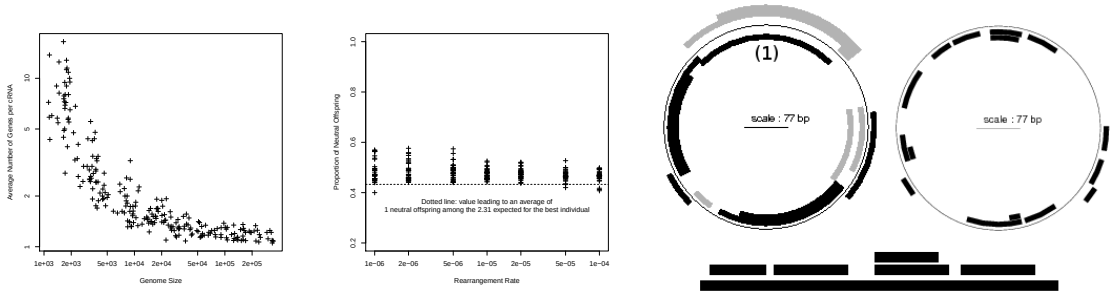


(a) (b) (c)

Figure A.4 – Évolution des paramètres structuraux des génomes pour une exécution « typique » de aevol. **(a)** Évolution de la taille du génome. **(b)** Évolution du nombre de séquences codantes (nombre de gènes). **(c)** Évolution de la taille moyenne des séquences codantes. D'après (Knibbe, 2006).

4 Résultats

Les différentes expériences que nous avons menées avec le modèle Aevol nous ont permis d'apporter des éléments de réponse à plusieurs questions ouvertes en biologie évolutive. En faisant varier les paramètres du modèle, nous avons observé de grandes variations dans l'organisation des génomes des individus. Nous avons ainsi constaté que, dans un environnement identique, une population évoluant avec un taux de mutations fort donnait naissance à des génomes beaucoup plus courts et compacts qu'une population sujette à des taux de mutations plus faibles (Knibbe et al., 2007a). Ce phénomène était déjà connu en ce qui concerne la quantité de séquences codantes sous la dénomination d'« error threshold » (Eigen, 1971; Ochoa, 2006) ou de fardeau mutationnel (Lynch, 2006), mais son extension à la quantité de non-codant constitue un résultat majeur du modèle.



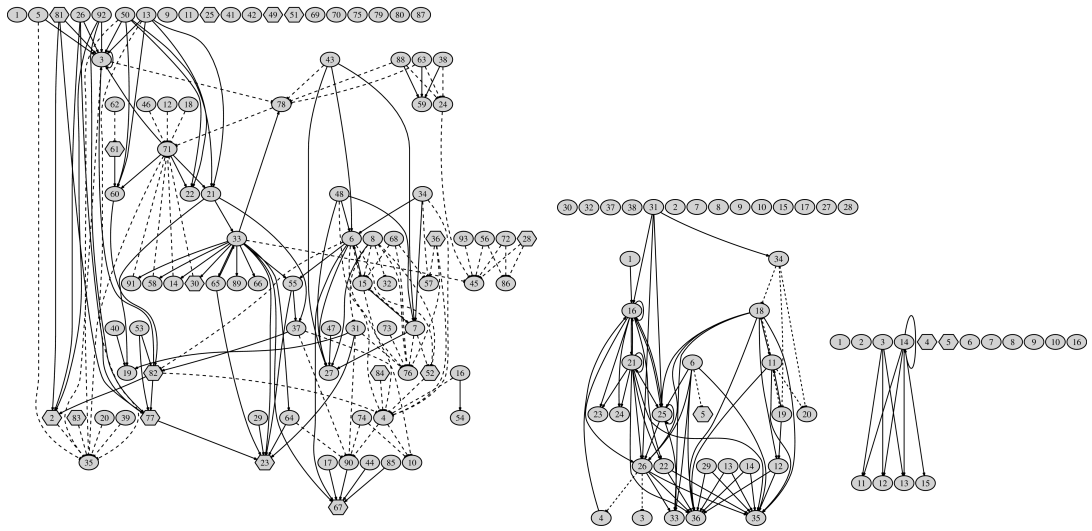
(a) (b) (c)

Figure A.5 – **(a)** Nombre de gènes par ARN codant (contenant au moins un gène) du meilleur individu de chaque population après 50 000 générations. **(b)** Proportion de descendants non neutres du meilleur individu de chaque population après 50 000 générations. **(c)** Génome du meilleur individu de la génération 50 000 d'une simulation typique avec des taux de mutations et de réarrangements élevés (1.10^{-4}). À gauche : ARNs (codants en noir, non-codants en gris). À droite : gènes. En bas : zoom sur l'opéron (1) avec ses 5 gènes.

Le modèle Aevol nous a également permis d'étudier l'influence des taux de réarrangements sur l'organisation de la transcription. Une analyse de la structure des ARNs d'organismes ayant évolué dans Aevol montre que les variations de taille des génomes s'accompagnent de profondes différences dans la façon dont ils sont transcrits. Les génomes les plus longs présentent de très nombreux ARNs non-codants, leurs ARNs codants étant courts et ne codant généralement que pour une seule protéine. Les génomes courts, quant à eux, sont généralement transcrits en des ARNs beaucoup plus longs codant chacun pour plusieurs protéines, formant ainsi des *opérons* (figure A.5). L'origine évolutive des opérons dans les génomes réels est une question ouverte en biologie (Lawrence, 1999).
Nous avons constaté qu'il existait un seuil de taux de réarrangements au-delà duquel les opérons deviennent la règle plutôt que l'exception. Cet effet de seuil est en fait le résultat de la combinaison de deux pressions antagonistes. Selon le phénomène d'*error threshold*, seuls les génomes courts peuvent être transmis fidèlement lorsque le niveau de variations génétiques est élevé. Par ailleurs, la sélection des individus les plus adaptés à l'environnement tend ici à favoriser ceux ayant beaucoup de gènes. La conjonction de ces deux pressions résulte ainsi en une pression vers la compaction des génomes et, *in fine*, en la formation d'opérons (Parsons et al., 2010b).

En conduisant des expériences similaires avec le modèle R-Aevol, une extension de Aevol dans laquelle un processus de régulation explicite du niveau d'expression des gènes a été introduit, nous avons pu obtenir des résultats très intéressants quant à la complexité des réseaux de régulation. Après avoir laissé évoluer des population d'organismes artificiels dans un environnement stable ne nécessitant aucune régulation, nous avons constaté de grandes différences dans la complexité des réseaux de régulation obtenus. Ainsi, les organismes ayant évolué avec de forts taux de mutations et de réarrangements, en plus d'avoir des génomes courts et ne comportant que peu de gènes, présentent des réseaux

de régulation très peu connectés. À l'inverse, les organismes n'ayant été exposés qu'à de faibles taux de mutations et de réarrangements présentent de longs génomes comportant beaucoup de gènes ainsi que des réseaux de régulations densément connectés (figure A.6). Il apparaît en fait que la complexité du réseau de régulation découle directement de la pression de second ordre vers un niveau spécifique de variabilité mutationnelle du phénotype qui est exercée sur les génomes. En effet, quand le nombre de gènes augmente, le nombre de promoteurs augmente également, résultant en une croissance super-linéaire du nombre d'associations possibles entre gènes et promoteurs. Nous avons ainsi pu montrer que, au moins dans le modèle, le taux de réarrangements est un déterminant majeur de l'évolution de la complexité des réseaux de régulation (Beslon et al., 2010b).



(a) (b)

Figure A.6 – (dans Beslon et al., 2010b). Réseaux de régulation du meilleur organisme à la dernière génération de simulations représentatives avec des taux de mutations/réarrangements respectivement bas **(gauche)**, modérés **(centre)** et hauts **(droite)**.

Étant donné le rôle majeur des réarrangements chromosomiques dans la structuration des génomes au travers de la pression de second-ordre que nous avons identifiée, une modélisation plus fine des mécanismes de réarrangements était nécessaire pour mieux tenir compte de leurs spécificités et en particulier de leur sensibilité aux similarités entre séquences.

Ici, nous avons modifié le modèle pour y introduire une sensibilité aux alignements entre séquences dans les mécanismes de réarrangements chromosomiques. Nous avons développé un algorithme de recherche d'alignements adapté au contexte de l'évolution expérimentale *in silico* qui permet de rendre compte de la plus grande probabilité qu'un réarrangement ait lieu entre des séquences similaires (réarrangements homologues) qu'entre des séquences qui diffèrent grandement (réarrangements non homologues).

En utilisant cette extension du modèle, nous avons montré que, si les réarrangements chromosomiques sont dangereux, ils sont également nécessaires pour que l'évolution soit

efficace. Ici, la plupart des réarrangements fixés sont des réarrangements homologues. Cependant, ces réarrangements nécessitant la présence de séquences répétées pour avoir lieu, celles-ci doivent être créées (par duplication) au moins aussi vite qu'elles ne sont dégradées (principalement par des mutations locales). Nous avons ainsi pu identifier une interaction complexe entre les réarrangements homologues et non homologues, pouvant créer des séquences répétées, et les mutations locales, pouvant les dégrader. Au final, il semble que les conditions dans lesquelles l'évolvabilité est maximale corresponde à des taux de réarrangements non homologues faibles combinés à des taux de mutation eux aussi faibles, permettant ainsi de maintenir un nombre limité de séquences répétées qui favorisent l'évolvabilité en constituant un bon substrat pour les réarrangements homologues.

Les transferts horizontaux jouent un rôle majeur dans l'évolution des bactéries en permettant de solutionner les problèmes d'interférence clonale mais aussi en fournissant un moyen de casser les déséquilibres de liaison qui pourraient apparaitre. Les déséquilibres de liaisons étant souvent cités comme condition nécessaire à l'action de pressions de second ordre, nous avons introduit dans le modèle un mécanisme de transfert horizontal biologiquement plausible permettant de casser ces déséquilibres en favorisant la recombinaison allélique. Les résultats obtenus avec cette nouvelle extension du modèle reproduisent très précisément les résultats précédemment obtenus avec le modèle "classique", ce qui confirme que, au moins dans le modèle, le contrôle du niveau de variabilité mutationnelle du phénotype est distribué sur l'ensemble du génome, déterminé par la structure du génome elle-même.

L'utilisation de Aevol nous permet de montrer les relations qui existent entre tous ces résultats. Ils traduisent en effet la nécessité, pour un organisme en évolution, de maintenir un équilibre entre robustesse et évolvabilité, entre capitalisation de l'acquis et exploration de nouvelles solutions. Ainsi, lorsque l'on analyse les meilleurs individus d'une population évoluée, on constate qu'ils partagent tous une caractéristique commune : leur nombre moyen de descendants neutres $F_\nu W$ (produit de la probabilité de reproduction neutre [1] F_ν et du nombre de descendants W – voir figure A.5) est juste supérieur à 1.
Aevol a donc montré que les génomes ne sont pas façonnés uniquement par des pressions directes sur la fitness ou par des biais mutationnels. Ils sont aussi profondément structurés par des pressions indirectes (pressions de second ordre) dont celle pour atteindre un bon compromis entre exploration et exploitation. La compacité du génome est un levier d'ajustement de ce compromis car les génomes présentant plus de gènes et plus de non-codant subissent plus de réarrangements pouvant impacter la fitness (Knibbe et al., 2007a).

5 Conclusion

En faisant évoluer, de façon réaliste, des organismes virtuels, Aevol nous permet d'étudier les mécanismes évolutifs. Aevol permet de retrouver de nombreuses caractéristiques struc-

1. Cette probabilité peut être obtenue expérimentalement en effectuant 10 000 reproductions de l'individu et en comptant le nombre de descendants ayant la même fitness que le progéniteur.

turelles des génomes d'organismes réels en faisant varier des paramètres tels que les taux de réarrangements ou la taille de la population. Il nous permet donc de proposer aux biologistes des hypothèses alternatives pouvant expliquer ces phénomènes. Le modèle Aevol peut donc être considéré comme un générateur d'hypothèses pour expliquer l'évolution de l'organisation des génomes.
D'un point de vue biologique, le modèle a vraisemblablement encore beaucoup à nous apprendre, nous projetons par exemple de mener des expériences parallèles sur le modèle et sur des organismes réels pour étudier l'évolution de la pathogénicité chez certaines bactéries.
D'un point de vue informatique, l'identification de pressions de second ordre telles que nous avons pu les observer dans nos simulations pourrait ouvrir de nouvelles voies dans le domaine de l'optimisation par algorithmes génétiques. D'autre part, les données produites par le modèles peuvent servir de banc d'essai pour des algorithmes de découverte de connaissances (Beslon et al., 2010a).