Importance of the rearrangement rates on the organization of genome transcription

David P. Parsons[1,3], Carole Knibbe[2,3] and Guillaume Beslon[1,3]

[1]Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, F-69621, France
[2]Université de Lyon, CNRS, Université Lyon 1, LIRIS, UMR5205, F-69622, France
[3]IXXI, Institut Rhône-Alpin des Systèmes Complexes, Lyon, F-69007, France
guillaume.beslon@liris.cnrs.fr

Abstract

The organization of genomes shows striking differences among the different life forms. These differences come along with important variations in the way genomes are transcribed, operon structures being frequent in short genomes and the exception in large ones, while ncRNAs are frequent in large genomes but rare in short ones. Here, we use the digital genetics model "aevol" to explore the influence of the mutation rates on these structures, showing that their diversity can be accurately reproduced when varying the rearrangement rate. This result points us to the mutational burden hypothesis as one of the main explanation. In this view, a specific level of mutational robustness indirectly leads to genome and transcriptome streamlining.

Introduction

Genome organization is well known to be very different throughout the different domains of life. On one extreme, viral genomes can be as short as 400 base-pairs long (Gago et al., 2009) and are usually very dense, with nearly no non-coding sequences and a lot of overlapping genes, although some exceptions were reported (Raoult et al., 2004). Eukaryotic multicellular organisms on the other extreme, have very long genomes (billions of base-pairs), a huge proportion of which is composed of non-coding sequences. These differences come along with variations in the way the genome is transcribed: On the one hand, short genomes, that are almost entirely transcribed, are commonly transcribed into long RNAs that can contain several genes. 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 microRNAs 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 organisation of transcripts and their relative importance remain open questions. Most efforts in these matters have been focused in understanding the evolution of operon structures. Operons are very interesting RNA structures where several coding sequences (often functionally-related) are packed together on a single RNA. 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 RNA is beneficial because they share their regulation sites, which means that mutations on the promoter will preserve the relative expression levels of the gene products. According to this, 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.
- Finally, the mutational burden theory 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 mutation (or 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 mutation rates will face a pressure for making genomes denser (Lynch, 2006; Knibbe et al., 2007). In some cases, this densification may reach a point where transcribed regions can actually merge or where a transcribed region can contain several translated sequences thus composing an operon. In extreme situations, genes can even share a part of their sequence and

overlap. This 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, operons should then 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 mutation rates, the selection strength and the population size.

Each of these theories have 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 either *in vivo* or *in vitro* as the underlying processes are complex and act on a very long time scale. Comparative genomics approaches are a way to circumvent this difficulty. However, they are based upon the static snapshots of the contemporary sequences and have to *infer* their evolutionary past.

Artificial life and *in silico* simulations have shown to be very useful in such cases, providing us with insights into complex mechanisms and shedding light onto second-order pressures that would have been difficult to identify otherwise (Wilke et al., 2001; Adami, 2006; Misevic et al., 2006; Knibbe et al., 2007; Beslon et al., 2009). They provide a dynamic view of the evolutionary process in a reasonable time and with a near-to-absolute control over parameters. In this paper, we propose to investigate the organization of transcripts using a modelling-simulation approach.

Aevol: A digital genetics model

To study the evolution of genome structure, we have developed an integrated model, Aevol, that simulates the evolution of a population of N artificial organisms. Although a description of the model has already been published (see Knibbe et al. (2008) and its supp. mat.), we provide here an overview of the most important principles that are necessary to have a good understanding of the results presented here.

Overview

In Aevol, each artificial organism owns a genome whose structure is inspired by prokaryotic genomes. It is organized as a circular double-strand binary string containing a variable number of genes separated by non-coding sequences (figure 1). At the beginning of the run, all organisms are initialized with a same random sequence (of 5,000 base-pairs here) containing at least one gene. Genes are identified and decoded thanks to predefined signalling sequences and to an explicit transcription-translation process. Then, an abstract "folding" process gives rise to artificial "proteins" that are able to realize or deflect a particular range of abstract "biological functions". The interaction of all these proteins yields the set of functions the organism is able to perform, which will in turn be compared to an environmental target to determine how well-adapted this individual is.

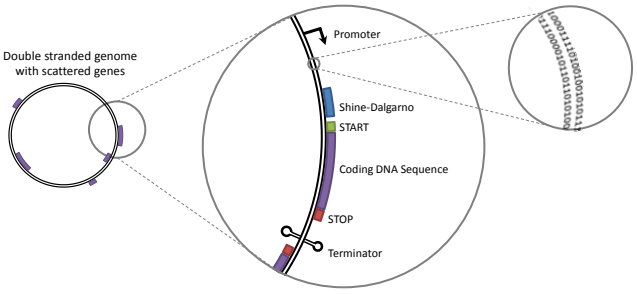


Figure 1: In Aevol, each individual owns a circular double-stranded binary genome upon which coding sequences are identified thanks to predefined signalling sequences: Promoters and terminators mark the boundaries of transcribed sequences and, inside these transcribed regions, coding sequences can exist between a START signal and an in-frame STOP codon (see figure 2 for the genetic code).

The best adapted individuals have higher chances of reproduction: At each generation, N new individuals are created by reproducing preferentially the best individuals of the parental generation which is then completely replaced. During the replication process, the chromosome can undergo different kinds of modifications: local mutations (single base substitutions, small insertions and small deletions), but also large chromosomal rearrangements (duplications, deletions, translocations and inversions).

From genotype to phenotype

The way a genotype is mapped to a phenotype in Aevol has been inspired by the prokaryotic transcription and translation processes. We defined a set of signalling sequences that enable us to identify the sequences that will be transcribed into RNAs and those that will be translated into proteins. Besides, a simple "folding" process was defined that allows us to interpret a protein's primary sequence as a set of "biological functions".

Transcription In prokaryotes, transcription initiates at particular sites, called promoters, where the RNA-polymerases recognize a consensus sequence to which they can bind and begin the RNA synthesis process. In Aevol, we defined a long consensus sequence, a promoter being a sequence whose Hamming distance d with this consensus is less than or equal to d_{max}. In the experiments presented here, the consensus was the 22-base-pairs (bp) sequence 0101011001110010010110 and up to d_{max} = 4 mismatches were allowed. This consensus sequence is long enough to

ensure that random, non-coding sequences have a low probability to become coding by a single mutation event. It is not a palindrome, meaning that a given promoter can initiate transcription on only one strand.

When a promoter is found, the transcription goes on until a terminator is reached. Terminators must be more frequent than promoters to limit the overlapping of transcribed sequences. Thus, if we had used a consensus sequence as for promoters, this sequence would have had to be very short. This would have forbidden this short motif to be present in any coding sequence, hence heavily constraining the evolutionary process. We therefore defined terminators as sequences that would be able to form a stem-loop structure, as the ρ-independent bacterial terminators do. In these experiments, the stem size was set to 4 and the loop size to 3, terminators thus had the following structure: $abcd***\bar{d}\bar{c}\bar{b}\bar{a}$, where $a, b, c, d = 0$ or 1.

The probability of a random 22-bp long sequence to be a promoter (*i.e.* of being at most 4 mismatches away from the consensus) is of roughly $1/460$, which means that the average distance between two promoters that can be expected in a random double-stranded sequence is of 230 bases. Terminators should be much more frequent: An 11-bp long sequence has a probability of $1/16$ to be a terminator.

The expression level e of an RNA is determined according to its promoter sequence. The closer the promoter is from the consensus, the higher the expression level: $e = 1 - \frac{d}{d_{max}+1}$. This modulation of the expression level models in a simplified way the basal interaction of the RNA polymerase with the promoter, without additional regulation. It provides duplicated genes with a way to reduce temporarily their phenotypic contribution while diverging toward other functions. It also induces a link of co-regulation between the coding sequences of a same transcribed region, which is a necessary property to test the coregulation hypothesis.

Translation Transcribed sequences (RNAs) do not necessarily result in a protein. The translation process of an RNA takes place when a Shine-Dalgarno-like sequence is found, followed, a few base-pairs away, by a START codon (see genetic code on figure 2). We thus defined the translation initiation signal as the motif $011011***000$. Whenever this signal is found, the following sequence is read three bases (one codon) at a time until the termination signal (the STOP codon 001) is found on the same reading frame. Each codon lying between the initiation and termination signals is translated into an abstract "Amino-Acid" using an artificial genetic code, therefore giving rise to the protein's primary sequence (figure 2).

As in real organisms, and because we read our genetic sequences three bases at a time, genes can be found on six different reading frames (three on each strand), giving the possibility for the organisms to evolve out-of-phase overlapping genes, which are commonly found in bacterial operons

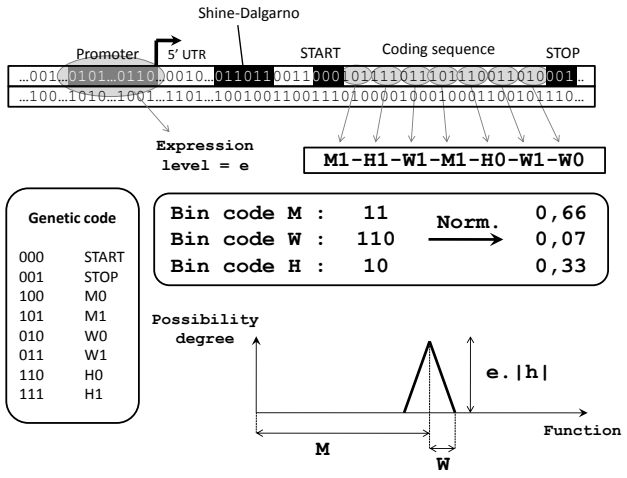


Figure 2: Overview of the transcription-translation-folding process in Aevol. Transcribed sequences are those that start with a promoter (consensus sequence) and end with a terminator sequence (hair-pin), not shown on the figure. Coding sequences (genes) are searched within the transcribed sequences; They begin with a Shine-Dalgarno-START sequence and end with a STOP codon. An artificial genetic code (right) is used to convert a gene into the primary sequence of the corresponding protein and a "folding process" enables us to compute the metabolic activity of this protein (functional abilities).

(Johnson and Chisholm, 2004; Palleja et al., 2008).

Protein "folding" and phenotype computation To model the activity of proteins and the resulting phenotype, we defined a simple "artificial chemistry" (Dittrich et al., 2001) that describes the organism's metabolism in a mathematical language. In our simplified artificial world, we assume that there is an abstract, one-dimensional space $\Omega = [0, 1]$ of possible metabolic processes (that is, in this model, a metabolic process is just a real number). In this "metabolic space", each protein is involved in a subset of processes (either realising it or preventing other proteins from realising it) which is described using the fuzzy set formalism: A given protein can be involved in a metabolic process with a possibility degree lying between 0 and 1. A protein is thus fully characterized by a mathematical function that associates a possibility degree to each metabolic process. For simplicity, we use piecewise-linear functions with a symmetric, triangular shape (figure 2). In this way, only three numbers are needed to characterize the metabolic activity of a protein: The position m ($m \in \Omega$) of the triangle on the axis, its half-width w and its height h (positive when realizing a function, negative when inhibiting it). This means that the protein contributes to the range $[m-w, m+w]$ of metabolic processes, with a preference for the processes closest to m

(for which the highest efficiency, h, is reached). Thus, various 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 this framework, each protein's primary sequence is decomposed into three interlaced binary subsequences that will in turn be interpreted as the values for the m, w and h parameters. For instance, the codon 010 (resp. 011) is translated into the single amino acid $W0$ (resp. $W1$), which means that it contributes to the value of w by adding a bit 0 (resp. 1) to its binary code. Small mutations in the coding sequence (substitutions, indels, possibly causing frame shifts) will change these parameters, resulting in a modification of the protein's metabolic activity.

Once all the proteins encoded on the genotype of the organism have been identified and characterized, their activities are combined into a fuzzy set representing the individual's phenotype P, using Lucasiewicz' fuzzy operators. This phenotype indicates to what extent the individual can realize each metabolic process in our abstract metabolic space.

Environment, adaptation and selection

In Aevol, the environment is represented by a phenotypic target: The fuzzy set E defined on Ω that represents the optimal degree of possibility for each "biological function". To evaluate an individual, we compare its phenotype P to the optimal phenotype E. The "metabolic error" g is computed as the geometric area between these two sets (figure 3). The lower the metabolic error, the better the individual. This measure penalizes both the under-realization and the over-realization of each function.

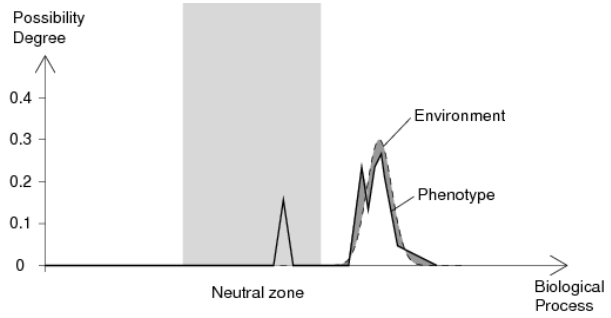


Figure 3: Measure of an individual adaptation. Dashed curve: Environmental target E. Solid curve: Phenotypic distribution P (resulting metabolic profile obtained after combining all the proteins). Dark grey filled area: Metabolic error g. The part of the phenotype that is located inside the neutral zone (light grey) is not considered as being part of the gap. This allows for the evolution of non-essential genes.

In the current version of Aevol, the population size is constant (here $N = 1,000$ individuals) and the population is entirely renewed at each generation. A probability of reproduction is assigned to each individual according to its metabolic error and a multinomial drawing determines the actual number of offsprings each individual will have. In the experiments presented here, we used an exponential ranking selection (Blickle and Thiele, 1996). The individuals are sorted by decreasing metabolic error so that the worst individual has rank $r = 1$ and the best $r = N$. The probability of reproduction of an individual is then given by $\frac{s-1}{s^N-1}s^{N-r}$, with $s = 0,998$ being the intensity of selection in all the experiments presented here.

Genetic operators

During their replication, genomes can undergo seven different kinds of modifications, three of which are local mutations (single nucleotide substitutions and insertions or deletions of 1 to 6 bp) and the four others, chromosomal rearrangements (duplications, deletions, translocations and inversions). The breakpoints for these rearrangements are randomly chosen on the chromosome.

Mutations and rearrangements affect the genome but do not necessarily have a phenotypic effect. For instance, a mutation that takes place in an untranscribed region will be completely neutral unless it creates a new promoter, which is reasonably rare given the size of the consensus sequence.

The rates at which each type of genetic modification i occurs (μ_i) are parameters of the model. They are defined as the per-base, per-replication probability of each type of modification to take place. Although horizontal transfer is possible in Aevol, we disabled it entirely in these experiments to avoid the assembly of operons due to the selfish operon effect.

Aevol is hence a digital genetics model in which the structure of the genome is free to evolve. It integrates major genetic features and mechanisms, introducing a transcription-translation level between the genetic and the phenotypic levels and allowing both local mutations and large chromosomal rearrangements. These particularities make Aevol a model that is particularly suited for the study of genome organization.

Results

The typical use of digital genetics models is very close to experimental evolution procedures (Elena and Lenski, 2003): Populations of organisms are initialized and left to evolve in controlled conditions. By observing the products and the dynamics of the evolutionary process in different conditions and by comparing them, we can unravel the direct or indirect pressures that constrain the structure of the organisms.

We let 147 populations of 1,000 individuals evolve during 20,000 generations in near identical conditions where the only changing parameters were the mutation rate and the

rearrangement rate (one common rate μ_m for the three different types of local mutations and one, μ_r, for the four types of rearrangements) for which values ranged from 1.10^{-6} to 1.10^{-4} per base-pair (7 rates tested). Each combination of mutation and rearrangement rates was tested with 3 independent seeds.

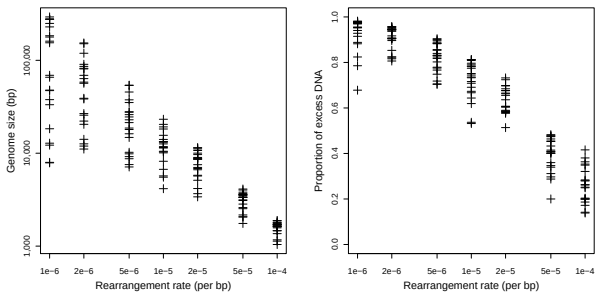
These populations evolved in identical environments composed of a single Gaussian curve placed on the right hand side of the metabolic axis (figure 3). The central zone of the axis was neutralized, meaning that the organisms receive no penalty for evolving proteins in that zone (even though they are of no use). This will enable us to test whether non-essential genes can be packed together with other genes in an operon structure.

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, operons should not be observed in such conditions. Operons that would arise nevertheless could be explained by either the co-regulation or the mutational burden hypotheses. The variations of mutation and rearrangement rates will enable us to test the mutational burden hypothesis, and the co-regulation theory can be tested by analysing the functional relatedness of genes organized in operons.

Evolution of the structure of the genome

During the evolutionary process, the organisms progressively acquire new genes 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. However, looking at the evolution of the size of the genome and the number of genes, we can see a clear trend for individuals evolving under lower rearrangement rates to have larger genomes containing both more genes and a greater proportion of non-coding sequences (figure 4). The rate of rearrangements is the major factor explaining the variability of genome compactness, the rate of small mutations has a much lower effect. Interestingly, the genome size stabilizes even though there is no direct cost for neither the replication of the genome nor for its expression.

As we have already shown, these effects are the consequence of the long-term selection of a specific level of mutational robustness (Knibbe et al., 2007). Indeed, we have estimated the fidelity of the replication for each of the 147 final best individuals, by a mutagenesis-like experiment: We let each of them reproduce 10,000 times and counted the number of offspring that had retained the ancestral fitness, in order to estimate the fraction of neutral offspring, F_ν. Figure 5 shows that in all cases, the genome had evolved in such a way that F_ν was greater than $1/2.31$. Thus, on the 2.31 offsprings expected for the best individual during the runs (given the selection intensity), at least 1 of them would retain the ancestral fitness, while the other ones would explore other phenotypes. This reflects the indirect selection of an appropriate trade-off between exploitation and exploration: under a high mutation rate per base-pair, the only way to reach a good trade-off is to keep the genome small. This phenomenon, known as an "error threshold" effect (Eigen, 1971), sets an upper bound to the total coding length, but also, here, on the non-coding length. Indeed, when rearrangements are taken into account, non-coding sequences are actually mutagenic for the genes they surround, because they provide breakpoints for large duplications or deletions (Knibbe et al., 2007).



(a) Genome size (b) Proportion of excess DNA

(c) Number of genes (d) Metabolic error

Figure 4: Genome size, proportion of excess DNA, number of genes and metabolic error for the best individual of each simulation after 20,000 generations. The fittest individuals are those with the lowest metabolic errors. Excess DNA includes here the intergenic DNA (between two coding RNAs) and the untranslated regions of the RNAs.

Evolution of the structure of transcripts

Looking more specifically at transcription-related features, our attention was drawn by the clear trend for higher rearrangement rates to favour long RNAs (figure 6(a)). The dynamics that leads to this lengthening of transcripts is very interesting: Indeed, as figure 7 shows, only the terminators seem to be gotten rid of during the whole evolutionary time,

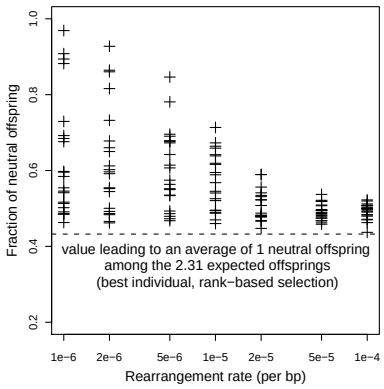
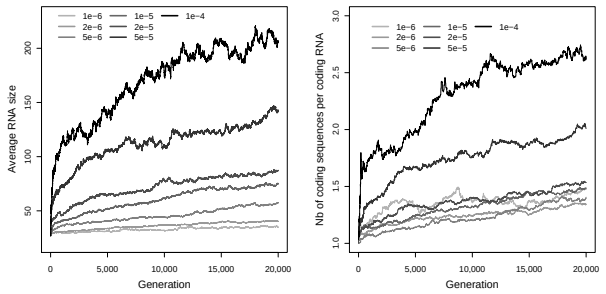


Figure 5: Fraction of neutral offspring estimated for the final best individual, after 20,000 generations of evolution.

the promoter density remaining stable after the first stage of evolution.

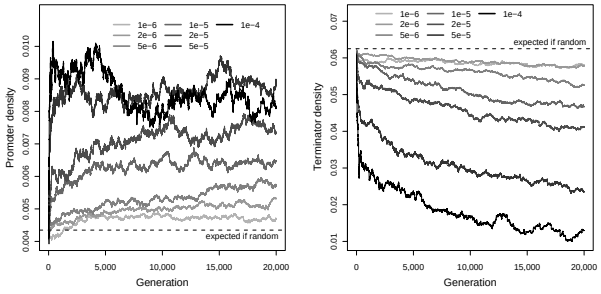


(a) RNA size (b) Number of genes per RNA.

Figure 6: Evolution of the average size of RNAs (regardless of whether they are coding or non-coding) and the average number of genes per coding RNA (RNAs containing at least one CDS). For clarity purpose, the data displayed here has been averaged over the different small mutation rates and seeds. Each line is hence the average value of the 21 simulations that were run under the same rearrangement rate.

Selection against terminators under high rearrangement rates leads to a lengthening of RNAs. But why are long RNAs selected for? What are the benefits of postponing transcription termination? The answer apparently resides in the packing of coding sequences: On average, RNAs belonging to organisms that evolved under high rearrangement rates own way more genes than those under low rates (figure 6(b)).

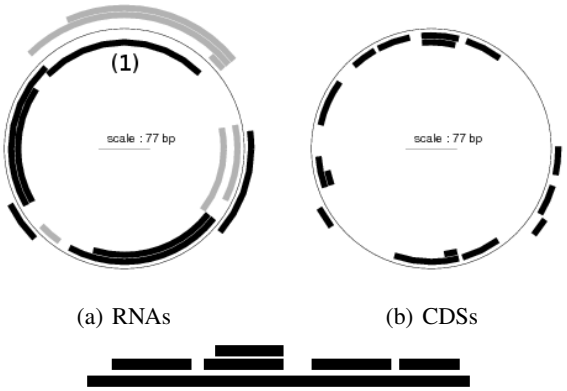
Figures 8 and 9 show the translation and transcription organization of the best individuals (after 20,000 generations) of 2 typical simulations with respectively high and low mutation and rearrangement rates. Under low rearrangement rates, almost every single CDS is transcribed by a different RNA. On the contrary, the individual that evolved under high



(a) Density of promoters (b) Density of terminators

Figure 7: Evolution of the average density of promoters (a) and terminators (b) for the different rearrangement rates. See figure 6 for details about data aggregated.

rearrangement rates has but one RNA containing only one gene, all the other transcripts carrying at least two. These figures also show a great difference regarding non-coding RNAs. At high mutation rates, a huge proportion of RNAs are ncRNAs whereas they become rare at high rearrangement rates, this reproduces what is observed in real organisms, eukaryotes having way more ncRNAs than prokaryotes have. Putting the focus on this aspect of our data, we found a clear scaling law between the rearrangement rate and the proportion of ncRNAs (data not shown). This scaling is a direct consequence of the proportion of non-coding sequences on the genome.



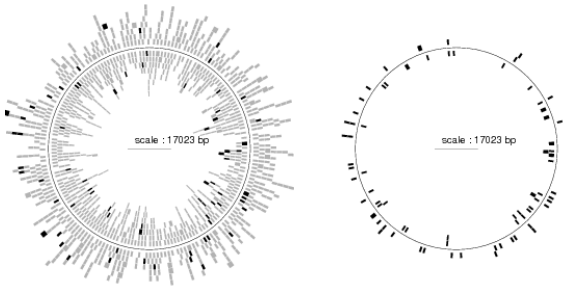
(a) RNAs (b) CDSs

(c) Zoom on operon (1) with its 5 genes

Figure 8: Genome of the best individual of generation 20,000 of a typical simulation with mutation and rearrangement rates of 1.10^{-4} per base-pair. In subfigure (a), coding RNAs are represented in black and ncRNAs in grey.

Discussion

In the experiments presented here, the organization of the genomes after 20,000 generations of evolution reproduces the whole range of genome organizations observed in real



(a) RNAs (b) CDSs

Figure 9: Genome of the best individual of generation 20,000 of a typical simulation with mutation and rearrangement rates of 1.10^{-6} per base-pair. In subfigure (a), coding RNAs are represented in black and ncRNAs in grey.

organisms. In our simulations, we observed a clear tendency for organisms having evolved under low rearrangement rates to have a eukaryote-like genome and for those under high mutation rates to resemble prokaryotic genomes.

Although a very small proportion of eukaryotic genomes is translated into proteins, a substantial fraction of these genomes is transcribed 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, ncRNAs have absolutely no function, yet they are very common when rearrangement rates are low. Interestingly, they are found at a proportion close to that which would be expected in a random sequence. Hence, it seems that ncRNAs are naturally present in intergenic regions making them available for acquiring new functions. It is tempting to suggest that these RNAs constitute a good substrate for the appearance of novel genes but this question will require a precise analysis of the dynamics of gene acquisition.

Another interesting feature we have observed is the emergence, under specific conditions (*i.e.* under high rearrangement rates), of operon structures.

Since operons appeared in a total absence of horizontal transfer, the selfish operon theory can easily be discarded as an explanation of the emergence of these operons. Indeed, horizontal transfer is a central and necessary feature of the selfish operon theory.

One of the remaining candidates to account for the emergence of the observed operons is the co-regulation model, under which hypothesis genomes should be more modular than expected at random. To compute the functional modularity of a genome, we conducted a pairwise comparison of the proportion of functionally related genes within operons and on the whole genome. Two genes were considered functionally related when they shared a subset of metabolic functions, *i.e.* when their corresponding phenotypic triangles overlapped. Given that the individuals evolved in a stable environment, no regulation is needed whatsoever. Modularity was shown to promote evolvability in the presence of inter-individual recombination (Pepper, 2000). However, here, reproduction was strictly clonal, which makes it difficult to imagine how the modularity of a genome could improve a lineage's evolutionary fate.

Yet, the results show a moderate tendency to pack functionally related genes together on the same operon: The proportion of pairs of functionally-related genes within operons was 1.26-fold higher (median value) than the same proportion on the whole genome. Although the effect is small, the ratio is significantly different from 1 (non parametric sign test, p-value = 7.10^{-4}).

These results do not allow us to conclude either in favor of or against the co-regulation theory and further experiments and analyses will be necessary to tackle this question.

According to the results presented in figure 6(b), there seems to be a threshold in the rearrangement rate above which operons become the rule rather than the exception. This is relevant when considered in the light of the mutational burden theory: As we have previously stated, the selection for a correct level of mutational robustness that was unravelled by Knibbe et al. (2007) leads to a strong pressure on the genome size. The higher the rearrangement rate, the smaller the genome must be to be transmitted faithfully to the offspring. Besides, the selection of the individuals that best fulfil the metabolic task (*i.e.* approximate the target) gives rise to a pressure for having many genes. Taken together, these two pressures result in the emergence of a composed pressure on the density of genes.

At medium rearrangement rates, the optimal gene density can be achieved by simply reducing the proportion of non-coding sequences, the coding sequences themselves remaining mostly unaffected. However, when the rates are really high, the amount of excess DNA (inter-RNA sequences, 3' and 5' UTRs, ncRNAs) shrinks to nearly nothing. At high rates, a further compaction can be done by several means such as making genes overlap (either on the same strand or on both strands) or getting rid of some of the transcription signals (promoters and terminators), hence merging consecutive RNAs into one single RNA (thus creating an operon).

We therefore expected to observe both overlapping genes and a lengthening of transcript length under high rearrangement rates. We indeed observed both of these phenomena (figures 8 and 6(a)) but were surprised by the dynamics leading to RNA lengthening: When the density of promoters appears to be stable over time, suggesting that they are not selected against, the density of terminators is constantly decreasing. Terminators fragment the genome, forbidding the sequences directly downstream from them (on both strands) to be translated, until a promoter is found. There is hence unmistakably a loss of gene density for each terminator on the genome. The solution that evolution found to efficiently pack genes together is then to limit this loss by decreasing the number of terminators on the genome, leading to a

lengthening of the average size of RNAs which in turn facilitates the emergence of operons.

Conclusion

In this paper, we have presented results that clearly reproduce features of genome organization that are observed in real organisms, in particular the structuration of genes in operons. The emergence of these operons specifically under high rearrangement rates points us to the mutational burden hypothesis, where a second-order pressure for a specific level of mutational robustness leads to genome streamlining. We now plan to conduct further experiments to investigate the role of horizontal transfer and how it interacts with this second-order pressure. We also plan to determine to what extent the co-regulation model can participate in the creation and maintenance of operon structures. Finally, we would like to analyse the role of non-coding RNAs in gene acquisition and to test whether they are innovation hot spots.

Acknowledgements

The authors would like to thank the BSMC group, who provides us with the computing resources and Fabien Chaudier for his invaluable help.

References

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

Beslon, G., Sanchez-Dehesa, Y., Parsons, D., Peña, J. M., and Knibbe, C. (2009). Scaling laws in digital organisms. In *Proc. Information Processing in Cells and Tissues IPCAT'09*, pages 111–114.

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

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

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

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.

Gago, S., Elena, S. F., Flores, R., and Sanjuan, R. (2009). Extremely high mutation rate of a hammerhead viroid. *Science*, 323(5919):1308.

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

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.

Johnson, Z. I. and Chisholm, S. W. (2004). Properties of overlapping genes are conserved across microbial genomes. *Genome Res.*, 14(11):2268–2272.

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.

Knibbe, C., Coulon, A., Mazet, O., Fayard, J.-M., and Beslon, G. (2007). 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. *Artif. Life*, 14(1):149–156.

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

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

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

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.

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

Palleja, A., Harrington, E., and Bork, P. (2008). Large gene overlaps in prokaryotic genomes: result of functional constraints or mispredictions? *BMC Genomics*, 9(1):335+.

Pepper, J. W. (2000). The evolution of modularity in genome architecture. In Maley, C. C. and Boudreau, E., editors, *Proceedings of the Evolvability Workshop at Alife VII*, Portland, USA.

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.

Raoult, D., Audic, S., Robert, C., Abergel, C., Renesto, P., Ogata, H., La Scola, B., Suzan, M., and Claverie, J.-M. (2004). The 1.2-megabase genome sequence of mimivirus. *Science*, 306(5700):1344–1350.

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.

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+.

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