Scaling Laws in Bacterial Genomes: A Side-Effect of Selection of Mutational Robustness?

G. Beslon*[,a,d], D. P. Parsons[a,d], Y. Sanchez-Dehesa[a,d], J.-M. Peña[c], C. Knibbe[b,d]

[a] *Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, F-69621, France*
[b] *Université de Lyon, CNRS, Université Lyon 1, LIRIS, UMR5205, F-69622, France*
[c] *DATSI, Universidad Politecnica de Madrid, Spain*
[d] *Institut Rhône-Alpin des Systèmes Complexes (IXXI), Lyon, F-69007, France*

Abstract

In the past few years, numerous research projects have focused on identifying and understanding scaling properties in the gene content of prokaryotes genomes and the intricacy of their regulation networks. Yet, and despite the increasing amount of data available, the origins of these scalings remain an open question. The RAevol model, a digital genetics model, provides us with an insight into the mechanisms involved in an evolutionary process. The results we present here show that (*i*) our model reproduces qualitatively these scaling laws and that (*ii*) these laws are not due to differences in lifestyles but to differences in the spontaneous rates of mutations and rearrangements. We argue that this is due to an indirect selective pressure for robustness that constrains the genome size.

Key words: Modelling, Scaling laws, Gene content, Transcription Factors, Mutational robustness, Evolvability

1. Introduction

Despite the huge diversity of living beings – from the smallest life forms to the biggest trees or mammals – some allometric ratios have been shown to be remarkably conserved among the living kingdom. For instance, it has been shown that various physiological characteristics of all organisms scale with their body mass and follow simple power-law behaviors whose exponents are multiples of $\frac{1}{4}$ (West et al., 2002). These scaling laws may reveal some fundamental principles of life, typically the necessity, for all organisms, to distribute energy and nutrients efficiently within their whole body (West and Brown, 2005).

At the molecular level, the ever-increasing number of sequenced genomes allows large comparative analysis. This analysis has revealed that several molecular traits also follow characteristic scaling laws. For instance, the genome size has been shown to scale as a power-law of the spontaneous mutation rate in DNA-based microbes (Drake, 1991; Drake et al., 1998). More recently, different genomic properties have been shown to follow power-law distributions (Luscombe et al., 2002; Koonin et al., 2002).

In prokaryotes, 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). When the sequenced bacterial genomes are considered globally, the diversity of genomic structure in prokaryotes is even more striking. Through the analysis of the annotated sequences, it was shown that the number of genes in each functional category scales as a power-law of the total number of genes in the genome and that the exponent of this law depends 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 (van Nimwegen, 2003; Molina and van Nimwegen, 2008). Moreover, this increase is also correlated with the size of the genome (Konstantinidis and Tiedje, 2004). These results suggest that the intricacy of regulation networks grows faster than the size of the network itself.

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 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

*Corresponding author: guillaume.beslon@liris.cnrs.fr

these scaling laws result from a selection process linked to bacterial lifestyle: Complex environments would require the coordination of multiple metabolic pathways (Cases et al., 2003). Alternatively, it has been 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 "molecular allometric laws" result from selective constraints (e.g., selection for short genomes or integrated networks), from the intrinsic dynamics of the evolutionary process or from any other mechanism still to be revealed (Molina and van Nimwegen, 2009).

In order to explore the evolutionary pressures on the genomic and transcriptomic structures and their dependence on external conditions (e.g., environmental conditions, population size, selection strength, mutation rates), an interesting approach is to use digital genetic models (Adami, 2006) where a finite population of virtual organisms is explicitly simulated in a virtual environment. These "organisms" are complex enough to be analysed in terms of molecular structure but they are also simple enough to allow for the computation of a fitness value, based on their genetic sequences and on the virtual environment. It is hence possible to implement a selection procedure. In such models, the evolutionary forces are precisely tuned and it is possible to test experimentally how they shape the structure of the organisms. Digital genetics have already shown that darwinian evolution can have counter-intuitive effects, due to indirect selective pressures. For example, it was shown that the long-term survival of a lineage not only depends on its fitness, but also on its mutational robustness (Wilke et al., 2001).

In this paper, we propose an integrated model of the evolution of regulatory networks, where the network level is not considered on its own but as a key layer between the genome sequence (where the mutations occur) and the phenotype (on which selection acts). We present our first large campaign of *in silico* experimental evolution with this model. Our results show that the model reproduces some known allometric laws, enabling us to propose hypotheses regarding their origin.

2. RAevol in a nutshell

To study the evolution of the structure of genomes and gene networks, we have developed an integrated model, RAevol (Regulatory-Aevol). This model extends the Aevol model (**A**rtificial **evol**ution), previously developed in our team to study robustness and evolvability in artificial organisms (Knibbe et al., 2007a,b, 2008). We provide here an overview of the Raevol model. A detailed description of the model is available in the Methods section.

In both Aevol and RAevol, 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). A set of predefined signaling sequences (promoters, terminators, Shine-Dalgarno-like sequences, start and stop codons) allows us to detect the coding sequences. These coding sequences are translated into abstract "proteins" that interact with one another and produce a phenotype that can be more or less well-adapted to the environment.

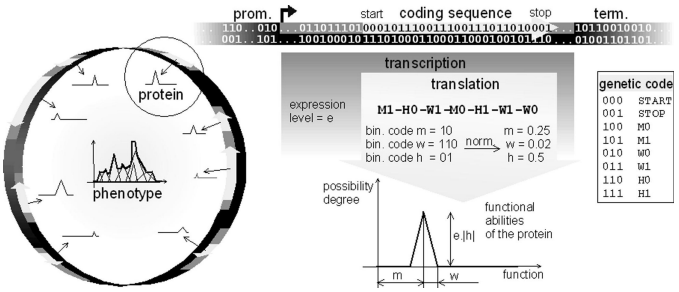


Figure 1: Overview of the transcription-translation-folding process in Aevol and Raevol. The genome is a circular, double-stranded, binary sequence (left and top). Transcribed sequences are those that start with a promoter consensus sequence and end with a terminator sequence. Coding sequences (genes) 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). In Aevol, the expression level e depends only on the sequence of the promoter and is constant throughout the lifetime of the artificial organism. In RAevol, e may vary over time due to the regulation activity of transcription factors. In both models, the expression level modulates the contribution of the protein (height of the triangle).

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 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 which is described using the fuzzy set formalism: A given protein can be involved in a metabolic process with a possibility degree comprised 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 1). In this way, only three numbers are needed to characterize the metabolic activity of a protein: The position m of the triangle on the axis, its half-width w and its height h. 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 coding sequence is translated into a chain of abstract "Amino-Acids" (AA) using an artificial genetic code (shown in figure 1). This 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 AA $W0$ (resp. $W1$), which means that it contributes to the value of w by adding a bit 0 (resp. 1) to its binary code. Thus, small mutations in the coding sequence (substitutions, indels, possibly causing frameshifts) will change these parameters, and hence the metabolic activity of the protein.

In the RAevol model each protein may have a regulatory activity beside its metabolic activity: It can interact with promoter sequences, thus enhancing or inhibiting the transcription of other genes. To determine whether a protein can regulate a particular promoter, we test whether the AA-chain of the protein contains a small motif that can bind to a subsequence of this promoter. The set of motifs that can bind to a particular DNA subsequence is randomly determined once and for all at the beginning of the evolutionary run. Like in most bacteria, the sign of the regulation depends on whether the binding occurs before or after the position of the first transcribed nucleotide (Janga and Collado-Vides, 2007). The resulting transcription level is used to scale up or down both the metabolic activity (height of the triangle) and the regulatory activities of the protein. We call the proteins that actually have a regulation activity Transcription Factors (TFs). Note that proteins with no metabolic activity (null w or h) can nevertheless be TFs. In this case, they are called *pure* TFs.

Due to this regulatory process, the transcription levels of the genes (and hence the protein concentration levels) may vary during the lifetime of the organism. At each time t, the global metabolism is computed by combining all the protein curves scaled by their concentrations. The phenotype of an artificial organism is thus defined as the dynamic curve showing the degree of realization of each possible metabolic process at each time t. The fitness of the organism is then computed as the distance between the phenotypic curve and a pre-defined target curve (representing the metabolic functions needed to survive in the environment). The fittest organisms are allowed to replicate, with small mutations and large rearrangements (duplications, deletions, inversions, translocations) occurring at random locations during genome replication. Genome size, gene number and gene order are hence free to evolve. Rearrangements can also modify the topology of the network (duplication or deletion of genes or promoter regions). Small mutations in coding sequences or in promoters can also affect the DNA-protein bindings and hence the wiring of the network.

3. Results

The typical use of digital genetics models is quite close to experimental evolution procedures (Elena and Lenski, 2003): Populations of organisms are initialized and left to evolve in controlled conditions (i.e., controlled parameters). By observing the products 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.

Eventually, our objective is to use RAevol 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 environment variations...). RAevol makes it possible to evolve digital organisms in demanding environments where they must react to external signals. However, we first wanted to check whether the organisms would evolve regulation networks in simple, steady, environments. Thus, we let the organisms evolve in a constant environment: 18 different populations of 1000 organisms evolved under 6 different mutation rates u (from 5.10^{-6} to 2.10^{-4} – defined as the per-nucleotide probability of a small mutation or a rearrangement occuring during replication), the selective pressure being exactly the same for all the experiments.

During the evolutionary process, the organisms progressively acquire new genes and connect them in such a way that they fulfill the task they are selected for (figures 2, 3 and 4). All the simulations proceed qualitatively in a similar way, evolving quickly in the first stage of evolution (rapid gene acquisition) 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 lower mutation rates to have larger genomes (figure 3) containing more genes (figure 4).

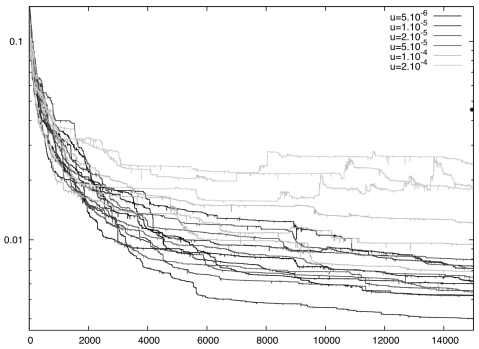


Figure 2: Evolution of the metabolic error of the best organisms of each simulation during 15000 generations (log scale). Whatever the mutation rate (except the highest), all organisms perform similarly.

We analyzed the structure of both the genomes and the regulation networks of the best organisms after 15000 generations. We found that all the features of the evolved organisms are influenced by the mutation rate: The organisms are clearly more complex when the mutation rate

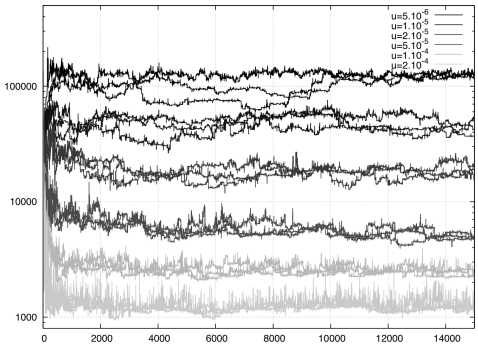


Figure 3: Evolution of the size of the genomes (in bp, log scale) of the best organisms during 15000 generations. The size of the genomes appears to be strongly dependent on the mutation rate u. Note that, in the model, genome size depends on both the number of genes and the size of non-coding sequences.

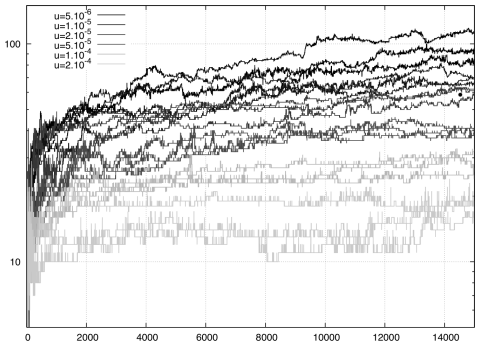
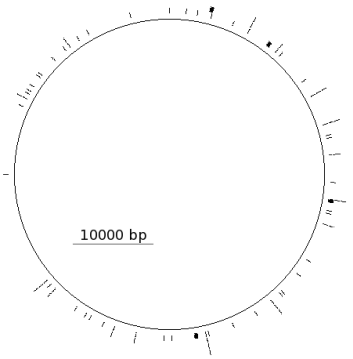
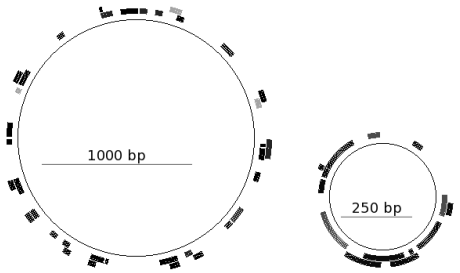


Figure 4: Evolution of the number of genes of the best organisms (log scale) during 15000 generations. After a short period of fast gene recruitment, the number of genes stabilizes. The number of genes in the organisms appears to be strongly dependent on the mutation rate u.



(a) A low mutation rate ($u = 5.10^{-6}$) leads to large genomes (here 120583 bp) with huge non-coding regions (here 97% of the genome).



(b) A medium mutation rate (left, $u = 5.10^{-5}$) leads to medium size genomes (here 4964 bp) with large non-coding regions (here 65% of the genome). A high mutation rate (right, $u = 2.10^{-4}$) leads to smaller genomes (1180 bp) with smaller non-coding regions (37%).

Figure 5: After 15000 generations, the genomes range from large ones (a) to intermediate and small (b) ones depending on the mutation rate u.

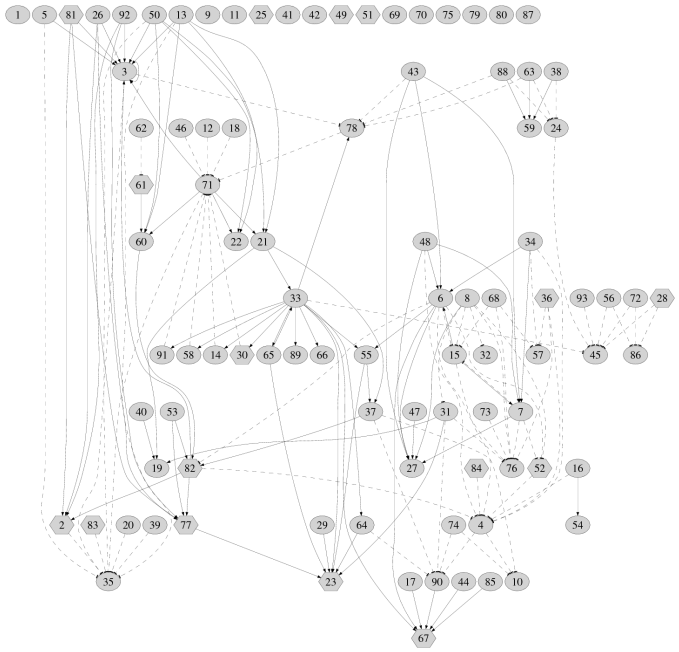
is low (figures 5 and 6) even though they all evolved in an identical and steady environment.

These results confirm the ones we had previously obtained with Aevol: The total coding length is influenced by the mutation rate and, much more surprisingly, the amount of non-coding sequences is also regulated (figure 7). With RAevol, we observe that the genetic network scales as well: The size and complexity of the network are clearly correlated with the mutation rate. In the simulations presented here, the environment is steady during the lifetime of the organisms. Thus, there is no direct pressure to evolve a regulatory network at all. Despite this, the lower the mutation rate, the more complex the evolved network. Both the number of genes and the number of TFs are inversely correlated with the mutation rate (figure 8). But as the mutation rate decreases, the number of TFs increases faster than the number of genes. This trend is even clearer in our runs if we consider the pure TFs (proteins with a regulatory activity but no contribution to the metabolism, figure 9).

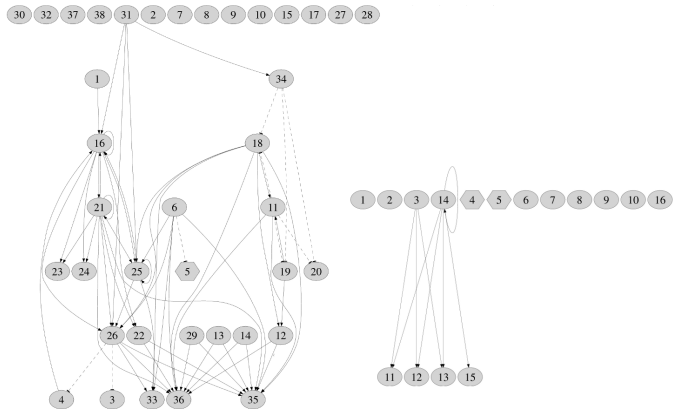
4. Discussion

As figures 10 and 11 show, our experiments with RAevol 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). Small genomes with few genes only have a very basic regulation activity while large ones develop complex regulation networks with many genes. Both the number of genes having a metabolic activity and the number of genes having a regulatory activity scale as power-laws of the total gene number, but when the former scales with an exponent below 1, the latter shows a super-linear scaling (figure 10).

In our experiments, all organisms evolved in the same – simple – environment. Thus, environmental conditions cannot have caused the scaling of the genetic complexity here. The only difference between our organisms was the mutation rate u that ranged from a very high one ($u = 2.10^{-4}$ mutations per bp per replication) to a low one ($u = 5.10^{-6}$ mutations per bp per replication). As figures 7 to 9 show, the mutation rate is the crucial factor determining the organisms' complexity. This is what we observed with the Aevol model in which proteins had no regulatory activity. We showed that this scaling was the consequence of an indirect selection of lineages whose genomic structure



(a) Low mutation rate ($u = 5.10^{-6}$) leads to high complexity (here 93 genes and 73 TFs, 13 of which being pure TFs).



(b) Medium mutation rate (left, $u = 5.10^{-5}$) leads to medium complexity (here 38 genes and 18 TFs). High mutation rate (right, $u = 2.10^{-4}$) leads to low complexity (16 genes and 2 TFs).

Figure 6: After 15000 generations, the complexity of the gene networks ranges from a high connectivity (a) to mild and low (b) ones depending on the mutation rate u. Solid lines represent activation links while dashed lines represents negative links. Genes having a metabolic activity are represented by ellipses. Hexagons represent genes without any metabolic activity.

allows for an appropriate trade-off between robustness and evolvability (Knibbe et al., 2007a, 2008, 2007b). If the per-base mutation rate is high, large genomes with many genes cannot maintain their fitness due to the mutational load they undergo. Large non-coding sequences cannot be maintained either because they promote large chromosomal rearrangements that can affect some genes. On the contrary, if the mutation rate is low, large genomes can maintain themselves in the population and they can even outcompete the smaller ones, because they can fit the target more precisely with more genes, and because they are

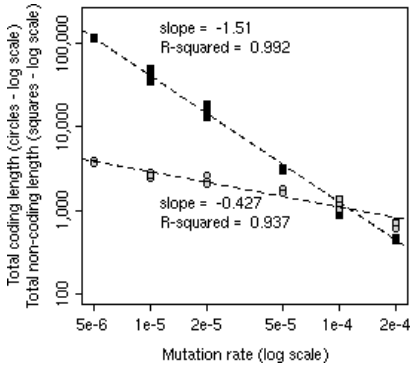


Figure 7: Size of the coding (gray circles) and non-coding (black squares) sequences for the best organisms of the 18 simulations at generation 15000 (log-log plot). Both values clearly scale with the mutation rate.

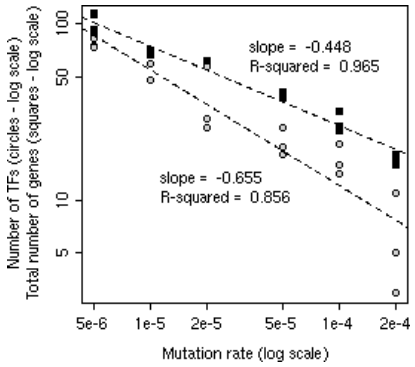


Figure 8: Total number of genes (black squares) and number of Transcription Factors (gray circles) for the best organisms of the 18 simulations at generation 15000 (log-log plot). Both values clearly scale with the mutation rate but the number of TFs grows faster than the number of genes.

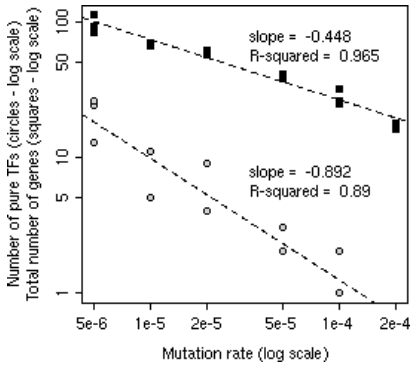


Figure 9: Total number of genes (black squares) and number of pure TFs (gray circles; Pure TFs are proteins having a regulation activity but no metabolic contribution) for the best organisms of the 18 simulations at generation 15000 (log-log plot). Both values clearly scale with the mutation rate but the number of pure TFs grows faster than the number of genes.

more likely to find a beneficial mutation. We showed for Aevol that this trade-off between robustness and evolvability manifested itself by the survival of the lineages whose expected fraction of neutral offpring F_ν (the expected fraction of offspring without mutation or only neutral ones at each reproduction) was close to $\frac{1}{W}$, where W is the number

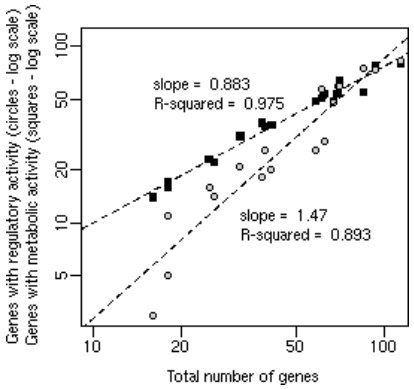


Figure 10: Number of genes involved in metabolism (black squares) and in the regulation process (gray circles) as a function of the total number of genes in the genome (best organisms of the 18 simulations at generation 15000; Log-log plot). Dash lines show power-law fits.

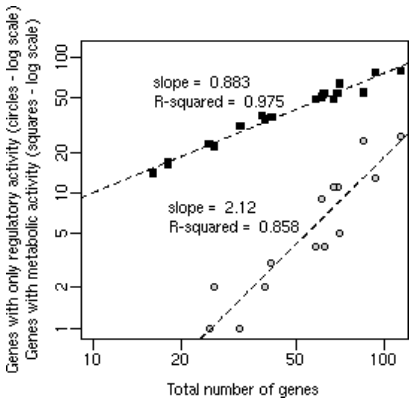


Figure 11: Number of genes involved in metabolism (black squares) and number of pure TFs (gray circles) as a function of the total number of genes in the genome (best organisms of the 18 simulations at generation 15000; Log-log plot). Dash lines show power-law fits.

of reproductive trials of the best individual (Knibbe et al., 2007a). In the experiments presented here, the evolved F_ν is again close to $\frac{1}{W}$ in most runs (figure 12). This suggests that the present results can also be explained by indirect pressures on the global mutational variability of the genome.

All the scaling laws observed in RAevol can derive from this pressure for robustness and from the scaling it imposes on the number of genes. Indeed, as the number of genes increases, the number of promoters also grows (possibly a little slower because of operon structures). Thus, the number of putative regulatory gene-promoter associations grows quadratically. Since, in the model, the regulatory activity is computed through a combinatorial algorithm that associates protein primary sequences with promoter sequences (see Methods), a linear increase in the number of promoters leads, for a protein with a regulatory motif, to a linear increase in the number of potential targets in the genome. As a consequence, a protein owning a regulatory motif has a higher probability of being a TF (number of actual targets in the genome greater or equal to 1) in a large genome than in a smaller one. Thus, RAevol appears as a

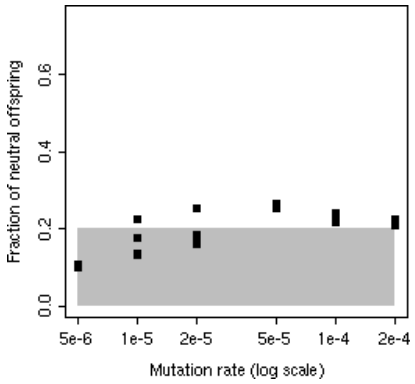


Figure 12: Evolved fraction F_ν of neutral offspring. F_ν was estimated by forcing the final best organism of each run to reproduce itself 10000 times (with the same mutation rate as during the run) and by counting the number of offspring that have the same fitness as their progenitor. The gray area represents organisms whose fraction of neutral offspring is lower than $\frac{1}{W}$.

null model in which links in the networks are added with an almost constant probability when the number of gene-promoter pairs increases. Consequently, in the model, the scaling of the number of genes (due to mutational robustness constraints) leads naturally 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 all their observations were based on genome annotation rather than on direct measures of the regulation networks' connectivity, it is difficult to assume such a combinatorial process to be at work. 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 number of TFs is also observed for *pure* TFs. Moreover, these proteins scale more than quadratically with the number of genes (figure 11). 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., genes 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. In this hypothesis, pure TFs could be directly selected. 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. We now plan to analyze the phylogeny of our organisms to study specifically the mechanisms that lead to the appearance and to the fixation of these "pure" regulators. We also plan to use KnockOut experiments to better understand their contribution to the dynamics of the regulation network.

To conclude, our results show that, at least in our model, the scaling laws reflect fundamental principles of bacterial evolution, i.e. the selection for an appropriate balance between robustness and evolvability (Lenski et al., 2006). Our simulations show that the pressure for complexification of the network can be indirect, unrelated to differences in the environment or the lifestyle: When facing identical environmental constraints, the organisms' 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 mutation rate here. Of course, this does not imply that, if faced with an environment of variable complexity and demand, organisms with the same mutation rate will have a similar structure. However, we can deduce from our results that the molecular complexity of the organism will be bound by the robustness constraint, meaning that the mutation rate will still be a major factor in determining organismal complexity.

5. Methods

5.1. Population initialization

Each population is seeded with 1000 asexual individuals with an identical genome. This initial genome is a random binary sequence of 5000 base pairs (bp) containing at least one coding sequence. Each run is seeded with a different initial genome.

5.2. Detection of transcribed regions

The transcription algorithm searches for promoters on each strand. Then, for each promoter, it follows the strand until it finds a terminator. This delimits the transcribed region. Note that several promoters can share the same terminator. In this case transcribed regions overlap.

Promoters are sequences similar to a pre-defined consensus. In the experiments presented here, the consensus sequence was 01010110011100100101 10 and $d \leq d_{\max} = 4$ mismatches were allowed. Terminators are sequences able to form a stem-loop structure, as the ρ-independent bacterial terminators do (here the stem size was set to 4 and the loop size to 3).

We assign a ground expression level β to the transcribed region depending on the similarity of the promoter with the consensus (Struhl, 1999): $\beta = 1 - \frac{d}{d_{\max}+1}$.

5.3. Detection of coding sequences and translation process

Once all transcribed regions have been localized, they are parsed to detect the initiation and termination signals of translation. These signals delimit the coding sequences. The initiation signal is the motif $011011***000$ (Shine-Dalgarno-like signal followed by a START codon, 000 here). The termination signal is the next STOP codon (001) on the same reading frame. Each time an initiation signal is found, the following positions are read three by three (codon by codon) until a stop codon is encountered. A transcribed region can contain several coding sequences (overlapping or not), meaning that operons are allowed.

Each coding sequence found inside a transcribed region is read triplet by triplet (codon by codon) and an artificial genetic code is used to translate it into a chain of artificial amino-acids. In this genetic code (shown in figure 1), there are 6 different amino-acids, grouped into three pairs (M_0/M_1, H_0/H_1 and W_0/W_1).

5.4. Metabolic activity of proteins

Let Ω be the abstract space of metabolic processes. To keep the model simple, Ω is one-dimensional space, more precisely a real interval: $\Omega = [a, b] \in \mathbb{R}$ (with $a = 0$ and $b = 1$ in the experiments presented here). Each protein i can contribute to (or inhibit) a fuzzy subset of Ω. This fuzzy subset is fully characterized by a mathematical function $f_i : \Omega = [a, b] \rightarrow [0, 1]$. This function is called a possibility distribution. It defines, for each metabolic process x the degree of possibility $f_i(x)$ with which the protein i can perform the process x. A metabolic process x belongs to the fuzzy set of a protein if $f_i(x) > 0$. We use piecewise-linear distributions with a symmetric triangular shape. Such distributions can be characterized by three parameters: The position m (mean) of the triangle on the axis, its height h and its half-width w. Hence a protein i can be involved in the metabolic processes ranging from $m_i - w_i$ to $m_i + w_i$, with a maximal degree of possibility for the process m_i. The fuzzy subset of the protein is thus the interval $]m_i - w_i, m_i + w_i[$.

In computational terms, the amino-acid chain of a protein is interpreted as three interlaced variable-length binary codes, giving the values of m_i, w_i and h_i respectively. To compute the value of m_i for example, we extract all M_0 and M_1 amino-acids found in the chain. They will form the Gray encoding of m (the Gray code is a binary numeral system where two successive values differ in only one bit). If the first M amino-acid of the chain is a M_0 (resp. a M_1), then the first bit of the Gray code of m_i is a 0 (resp. a 1), and so on. Thus, if the chain contains n amino-acids of type M, we get a Gray code of size n, which encodes an integer comprised between 0 and 2^{n-1}. A normalization enables us to bring the value of the parameter into

the allowed range, that is, $[a, b]$ for m. The same method is used to compute the values of w_i and h_i ($-1 \leq h_i \leq 1$ and $0 \leq w_i \leq w_{\max}$, $w_{\max} = 0.03$ here). If h_i is positive, the protein contributes to the metabolic processes.If h_i is negative, it impedes these processes. If h_i or $w_i = 0$ equals 0 it has no metabolic activity.

5.5. Regulatory activity of proteins

In Raevol, the transcription rate of a protein may vary throughout the lifetime of the artificial organism. It depends both on the intrinsic activity of the promoter (ground level, see above) and on the regulatory activity of the other proteins. Thus the concentration of a protein i is a function of time $c_i(t)$. This concentration is used to scale up or down the metabolic activity of the protein: The intrinsic distribution described above (triangle centred on m_i, of half-width w_i and of hight h_i) is multiplied by $c_i(t)$ at each time step. These scaled possibility distributions are those used to compute the phenotype at each time step (see below). This reflects the fact that a very efficient protein (high h_i) has actually no effect when it is not expressed. Similarly, the current concentration $c_i(t)$ of a protein also scales up or down the regulatory influence of the protein i on the other proteins at time t.

The possibility that a given protein will bind to a specific promoter is determined by a "value of affinity" between the amino-acid chain of the former and the genetic sequence of the latter. Small amino-acid motifs, that will henceforth be referred to as regulation domains, are able to bind to specific DNA subsequences with a given affinity. If a protein contains several regulation domains, its global affinity value over the promoter will be given by the best one among them. This value of affinity is used to determine the strength of the protein's influence on the transcriptional activity of the promoter it binds to. Like in most bacterial promoters, the nature of the regulation (activation or inhibition) depends on whether the binding occurs before (upstream) or after (downstream) the position of the first transcribed nucleotide (Janga and Collado-Vides, 2007). Thus, in RAevol, a promoter is composed of three DNA subsequences: The consensus sequence (where the RNA polymerase starts the transcription process) and its two flanking regions. When bound upstream, a protein enhances the transcriptional activity and, on the opposite, when bound downstream, it represses the activity of the polymerase, thus reducing the transcriptional activity.

The sequences that are able to interact with a specific DNA subsequence (thus constituting the possible regulation domains) are randomly determined at the beginning of the evolutionary run. In RAevol, regulation domains are small 5-Amino-Acid (AA) sequences that may have an affinity with 20-bp DNA sequences. To compute this affinity value, we align the regulation domain with the DNA sequence and compute the local affinity of each AA with the 4-bp subsequence it faces (figure 13). The motif will be able to bind the DNA sequence only if all five AA have

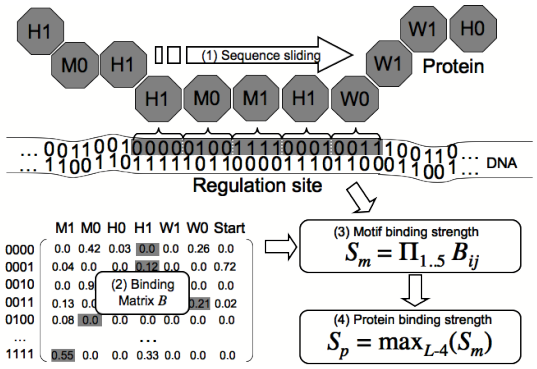


Figure 13: Computation of the binding between TFs and regulation sites. (1) The protein primary sequence slides in front of the 20-bp regulation site and all 5-AA motifs are tested. (2) For each AA-subsequence pair, the binding value B_{ij} is read in a Binding Matrix B (see main text for the initialization of the binding matrix). (3) The binding strength of the whole motif is the product of the five B_{ij} values and (4) the binding strength of the whole protein is the maximum strength over the $L - 4$ motifs it contains (L being the length of the proteins primary sequence).

strictly positive affinities with their corresponding DNA subsequences.

A binding matrix B is defined which contains the affinity of each amino-acid with each 4-bp sequence. Given our artificial chemistry principles, we have 7 possible amino-acids (START, M_0, M_1, H_0, H_1, W_0 and W_1) and $2^4 = 16$ 4-bp sequences. Thus, B is a 7×16 matrix. By choosing the initialization procedure of the regulatory matrix, we are able to choose the probability for a given motif to have a putative regulation activity. In all the experiments presented here, B was randomly initialized (uniform distribution in $[0, 1]$) and subsequently filled with 75% of null values. Thus, the probability that a given motif will bind to a specific DNA sequence of 20-bases long (length of the regulation sites in RAevol) is less than 0.1%. As a consequence, the probability that a 20-AA-long protein will be able to up-regulate (resp. down-regulate) a given promoter can be estimated at around 5% (probability to contain a motif that binds the promoter of the regulated gene).

The activity of a promoter depends on the sum of the activities of activators ($A_i(t) = \sum_j c_j(t) A_{ji}$) and on the sum of the activities of the inhibitors ($I_i(t) = \sum_j c_j(t) I_{ji}$), where A_{ji} (resp. I_{ji}) is the affinity of protein j on the enhancer of the promoter i (resp. on its operator) and $c_j(t)$ is the concentration of protein j at time t. When $A_i = I_i = 0$ (no regulation), the promoter has a ground activity β_i (Struhl, 1999). If $A_i > 0$ this activity increases progressively up to a maximum level. If $I_i > 0$, it decreases progressively to zero. The transcription rate e_i over time is then given by Hill-like functions:

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

where n and θ are constant coefficients that determine the shape of the Hill-function. In the simulations presented here, $n = 4$ and $\theta = 0.5$. 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 2). Thus, when a protein is regulated, its concentration is scaled up or down depending on its transcription rate.

$$\frac{\partial c_i}{\partial t} = e_i(t) - \phi c_i(t) \tag{2}$$

ϕ being a temporal scaling constant.

The transcription regulation in RAevol is a simplification of the real mechanisms of DNA-protein interaction. However, it catches the main mechanisms of genetic regulation while remaining computationally tractable. It also allows for proteins that perform a metabolic activity without any regulatory activity or, on the opposite, for proteins without any metabolic activity (i.e. $\int_0^1 |f(x)| = 0$) ta have a regulatory activity. We call "Transcription Factors" (TFs) the proteins that have a regulatory activity (regardless of their metabolic activity). Proteins having a regulation activity without contributing to the metabolism are called *pure* Transcription Factors).

5.6. Phenotype computation

Once all the proteins encoded on the genotype of the organism have been identified, the global phenotype can be computed by combining the whole set of proteins. We use the same formalism for the phenotype as for the proteins: The phenotype is the fuzzy subset of metabolic processes that the organism is able to perform. This fuzzy subset is described by a possibility distribution P indicating to what extent the organism is able to perform each process of Ω. The fuzzy logic framework provides us with logical operators to compute the complement, the union and the intersection of fuzzy subsets. Here, in logic terms, the global functional abilities of an individual are the metabolic processes that are enabled AND NOT disabled by the proteins of the organism.

$$P = (\cup_i (f_i | h_i > 0)) \cap \overline{(\cup_j (f_j | h_j < 0))} \tag{3}$$

Here, we use Lukasiewicz' fuzzy operators. For two proteins characterized by the distributions f_1 and f_2 respectively, Lukasiewicz' operations are defined as follows:

$$\begin{cases} \text{NOT:} & f_{\text{not}(1)}(x) = & 1 - f_1(x) \\ \text{OR:} & f_{1\cup 2}(x) = & \min(f_1(x) + f_2(x), 1) \\ \text{AND:} & f_{1\cap 2}(x) = & \max(f_1(x) + f_2(x) - 1, 0) \end{cases} \tag{4}$$

5.7. Fitness evaluation

Using our artificial chemistry, we are able to map a genotype to a phenotype, the latter being a dynamic function $P(t)$ which expresses the metabolism of the organism in the abstract functional space Ω. This enables us to evaluate each organism and to compute its "metabolic error" g in a given environment: The environment is described as a target (fuzzy) set of metabolic processes that have to be fulfilled by the cell in order to be able to reproduce. The metabolic error is computed as the area of the gap between both functions (figure 14). The lower the metabolic error, the higher the reproduction probability.

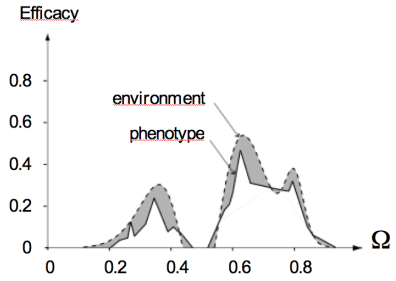


Figure 14: Measure of an individual adaptation. Dashed curve: Environmental distribution E. Solid curve: Phenotypic distribution P (resulting metabolic profile obtained after combining all proteins). Filled area: Metabolic error g.

Since the phenotype is a dynamic function, the environment may also be a dynamic function $E(t)$. Depending on the experiment one wants to do with the model, the metabolic error can be computed only once (e.g., after a transient period), at regular steps, during a time interval or after a particular environmental event. In this last case, the event can be sensed by the cell through "signaling molecule" which concentration may follow the environment variation. Here, the phenotype is computed during 20 time steps, the gap being computed at each time step during the second half. The metabolic error is then the mean of the 10 gap values.

5.8. Reproduction, mutations and rearrangements

In the current version of RAevol, the population size is constant ($N = 1000$ individuals here) and the population is completely renewed at each generation. At each generation, each individual is evaluated and a selection process is used to determine the number of offspring it will have. Then, all the selected organisms reproduce to create the next generation.

We use the "exponential ranking" selection scheme. At each generation, the individuals are sorted by decreasing metabolic error, such that the best individual has rank N. Then the probability of reproduction of the individual with rank r is $\frac{s-1}{s^N-1} s^{N-r}$, where $s \in]0, 1[$ tunes the intensity of the selection ($s = 0.995$ here). Finally, the actual numbers of reproductions are drawn by a multinomial drawing.

During their replication genomes can undergo seven different kinds of mutations, the first three being point mutations (switches and 1 to 6 bases indels) and the four others, large chromosomal rearrangements:

- Translocation: A randomly chosen segment of the genome is moved from its current position to a randomly chosen position.

- Inversion: A randomly chosen segment is inverted from one strand to the other and from one direction to the opposite one.
- Duplication: A randomly chosen segment is duplicated and reinserted at a randomly chosen position.
- Deletion: A randomly chosen segment is deleted.

Mutations affect the genome but can be neutral, for instance when they happen inside non-transcribed, non-coding regions. They can change the size of the genome, the number of genes or the functions of the proteins. Indirectly, they can modify the topology of the regulatory network, by either duplicating/deleting genes or promoter regions. Finally, they can modify the affinities between transcription factors and regulatory regions by changing either the promoter sequences or the regulation domain in the proteins' primary sequence.

The rate at which mutations occur, u (probability of mutation per base pair), is a parameter of the model. Here, in a given run, u was the same for all types of mutations. Six rates were tested: $u = 5.10^{-6}$, 10^{-5}, 2.10^{-5}, 5.10^{-5}, 10^{-4} and 2.10^{-4} per base pair. For each value, 3 independent runs were carried out.

Acknowledgment

The authors would like to thank Michael Parsons and Jean-Baptiste Rouquier for their help on the manuscript. The BSMC group provides us with the computing resources and we would like to warmly thank Fabien Chaudier for his invaluable help. This work has been funded by the french ANR MDCO Bingo2 2008-2010 project and the Spanish Ministry of Education (project number TIN2007-67148).

References

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

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

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

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

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

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., Crow, J. F., 1998. Rates of spontaneous mutation. Genetics 148 (4), 1667–1686.

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

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

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

Janga, S. C., Collado-Vides, J., 2007. Structure and evolution of gene regulatory networks in microbial genomes. Research in Microbiology 158 (10), 787–794.

Knibbe, C., Coulon, A., Mazet, O., Fayard, J.-M., Beslon, G., 2007a. A long-term evolutionary pressure on the amount of noncoding dna. Molecular Biology and Evolution 24 (10), 2344–2353.

Knibbe, C., Fayard, J.-M., 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.

Knibbe, C., Mazet, O., Chaudier, F., Fayard, J.-M., 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.

Konstantinidis, K. T., 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.

Koonin, E. V., Wolf, Y. I., Karev, G. P., 2002. The structure of the protein universe and genome evolution. Nature 420 (6912), 218–223.

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

Luscombe, N. M., Qian, J., Zhang, Z., Johnson, T., 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.

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

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

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

Reymond, N., Calevro, F., Viñuelas, J., Morin, N., Rahbé, Y., Febvay, G., Laugier, C., Douglas, A., Fayard, J.-M., 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.

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., 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.

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., 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.

West, G. B., Brown, J. H., 2005. The origin of allometric scaling laws in biology from genomes to ecosystems: towards a quantitative unifying theory of biological structure and organization. J Exp Biol 208 (Pt 9), 1575–1592.

West, G. B., Woodruff, W. H., Brown, J., 2002. Allometric scaling of metabolic rate from molecules and mitochondria to cells and mammals. Proc Natl Acad Sci USA 99 (supp. 1), 2473–2478.

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