BioSystems 102 (2010) 32–40

Contents lists available at ScienceDirect

BioSystems

journal homepage: www.elsevier.com/locate/biosystems

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 Politécnica de Madrid, Spain
[d] IXXI, Institut Rhône-Alpin des Systèmes Complexes, Lyon F-69007, France

ARTICLE INFO

Article history:
Received 4 May 2010
Accepted 15 July 2010

Keywords:
Modelling
Scaling laws
Gene content
Transcription Factors
Mutational robustness
Evolvability

ABSTRACT

In the past few years, numerous research projects have focused on identifying and understanding scaling properties in the gene content of prokaryote 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.

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 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 largescale 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

* Corresponding author at: Université de Lyon, CNRS, INRIA, INSA-Lyon, LIRIS, UMR5205, F-69621, France. Tel.: +33 4 72 43 84 87.
E-mail address: guillaume.beslon@liris.cnrs.fr (G. Beslon).

0303-2647/$ – see front matter © 2010 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.biosystems.2010.07.009

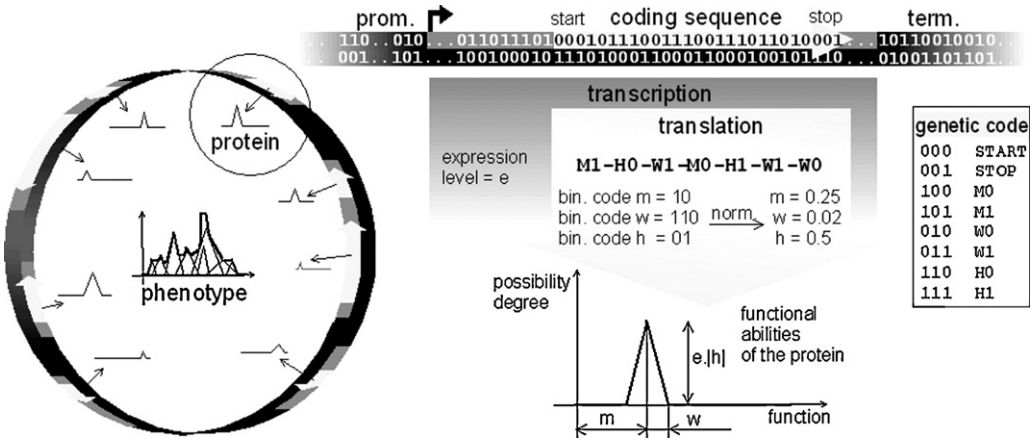

Fig. 1. Overview of the transcription-translation-folding process in Aevol and RAevol. The genome is a circular, double-stranded, binary sequence (left and top). Transcribed sequences are those that start with a promoter consensus sequence and end with a terminator sequence. Coding sequences (genes) 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. It is constant throughout the lifetime of the artificial organism and directly modulates the contribution of the protein (height of the triangle). In RAevol, e may vary over time due to the regulation activity of transcription factors. The expression level $e(t)$ and a degradation rate ϕ are then used to compute the protein concentration $c(t)$ which modulates its metabolic contribution. The height of the triangle representing the functional abilities of the protein then becomes $c(t)\cdot|h|$ (see Section 5 and Eq. (2)).

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 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 genetics models (Adami, 2006) where a finite population of virtual organisms is explicitly simulated in a virtual environment. These "organisms" are complex enough to be analyzed 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 has already shown that Darwinian evolution can have counter-intuitive effects, due to indirect selective pressures on variability. Indeed, since the mutational variability of the phenotype is partly under genetic control there can be a polymorphism in the level of variability in a population. Moreover, the variability level can influence the survival of lineages: those with inappropriate levels of variability can go to extinction due to a lack of robustness or evolvability – defined as the capacity of a lineage to generate adaptive heritable genotypic and phenotypic variation (Nehaniv, 2005). Thus there can be an indirect selective pressure on the factors that control the mutational variability of the phenotype: the mutation rate (Sniegowski et al., 2000), but also the properties of the genotype–phenotype map like modularity (Wagner and Altenberg, 1996).

Such indirect pressures are difficult to unravel in real organisms. Yet they can easily be studied using digital genetics experiments. For example, it was shown that, under high mutation rates, the indirect selection for mutational robustness can be strong enough to overcome the direct selection of immediate fitness, an effect called "survival of the flattest" (Wilke et al., 2001). It was also shown that a specific gene order can evolve by indirect selection of robustness against crossing-over (Pepper, 2003).

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 (Artificial evolution), 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 Section 5.

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 (Fig. 1). A set of pre-defined 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.