Parsimonious Modeling of Scaling Laws in Genomes and Transcriptomes

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

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

Abstract

We report here the use of Aevol, a software developed in our team to unravel the indirect selective pressures (i.e. pressures for robustness and/or evolvability) that act on the genome and transcriptome structures. Using Aevol, we have shown that these structures are under strong – although indirect – pressure due to the mutagenic effect of chromosomal rearrangements. Individuals undergoing high spontaneous rearrangement rates show more compact structures than individuals undergoing lower rates. This phenomenon concerns genome size and content (non-coding DNA, presence of operons, number of genes) as well as gene network (number of nodes and links) thus reproducing parsimoniously a large panel of known biological properties. The results reported here have been published in *Mol. Biol. Evol.* (Knibbe et al., 2007), *Biosystems* (Beslon et al., 2010) and Alife XII (Parsons et al., 2010).

Introduction

Largescale comparative analysis of sequenced genomes has revealed that several molecular traits 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). More recently, different genomic properties have been shown to follow power-law distributions (Luscombe et al., 2002). In prokaryotes for instance, it was shown that the number of genes in each functional category scales as a power-law of the total number of genes and that the exponent of this law depends on the functional role of the family: The number of transcription factors, in particular, scales quadratically with the total number of genes while metabolic genes scale linearly (van Nimwegen, 2003). This increase is also correlated with the size of the genome (Konstantinidis and Tiedje, 2004).

The origins of such scaling laws remain an open question. 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) or from the neutral dynamics of the evolutionary process.

An original approach to study the origins of genomic structures is to use *in silico* models of evolution. In such models, the evolutionary forces are precisely tuned and it is possible to test experimentally how they shape the organisms' structure. *In silico* evolution has already shown that darwinian evolution can have counter-intuitive effects, due to indirect selective pressures. For example, using the avida framework, Wilke et al. (2001) have shown that the long-term survival of a lineage not only depends on its fitness, but also on its mutational robustness. However, most digital genetic frameworks lack a precise description at the molecular level. That is why we have developed Aevol ("Artificial Evolution") and its extension R-Aevol ("Regulation in Aevol"). It specifically focuses on the molecular level in order to unravel the evolutionary pressures that act on genomes and transcriptomes. We report here the main results we got with Aevol. These results have been published in *Molecular Biology and Evolution* (Knibbe et al., 2007), *Biosystems* (Beslon et al., 2010) and Alife XII (Parsons et al., 2010). Aevol is freely available upon request from the authors.

The Aevol model

In Aevol, organisms own a circular, double-stranded genome of binary "nucleotides". Predefined signaling sequences as well as an artificial genetic code allow to detect the coding sequences and to translate them into abstract "proteins". We defined an artificial chemistry that describes the metabolism in a mathematical language: We assume that there is a one-dimensional space of all possible metabolic functions in which proteins are represented by a subset describing their metabolic contribution. This subset is described by parameters encoded in the coding sequence of the protein. Mutations in this sequence change these parameters, hence the metabolic activity of the protein.

In Aevol, the transcription rate of a given gene depends only on its own promoter sequence. In R-Aevol, proteins may have a regulatory activity besides their metabolic activity, thus being able to enhance or inhibit the transcription of other genes by binding to their promoters. The resulting transcription level is used to scale up or down both the metabolic and the regulatory activities of the protein. Due to this regulatory process, the transcription levels of the genes

vary during the organism life and so do the protein activities.

In Aevol as well as in R-Aevol, the global metabolism is computed by combining all the proteins' activities and the phenotype represents the degree of realization of each possible metabolic function. The fitness of the organism is then computed as the distance between the phenotype and a predefined target. The fittest organisms are allowed to replicate, with small mutations and large rearrangements (duplications, deletions, inversions, translocations) occurring randomly during the replication. Thus the genome size, gene number and gene order are free to evolve. In R-Aevol, mutations and rearrangements can also modify the regulatory network by either duplicating/deleting genes or promoter regions or by modifying their binding potentials.

Results

Digital genetics models are experimental models: Population of individuals evolve in different conditions and, by observing the genomic and transcriptomic structures of the evolved organisms, one then links the structures to the evolutionary conditions. Analysis of the lineages then enables to unravel the origins of the observed structures, ideally by discovering invariant properties in all simulations. In Aevol, we classically explore the influence of mutation rates, rearrangement rates and selection strength. The most striking results were obtained by exploring the influence of rearrangement rates on the different organization levels of the model:

I) By observing the genome length of evolved organisms, we observed a linear scaling between the rearrangement rate and the length of the non-coding sequences in Aevol's genomes. We have shown that this scaling is due to an indirect selective pressure acting on the non-coding sequences: Due to chromosomal rearrangements, non-coding sequences have a mutagenic effect on the surrounding genes. This long-term selective pressure offers a new explanation to variability of genome size and content (Knibbe et al., 2007).

II) By reproducing in R-Aevol the same experiment, we have shown that this pressure also acts at the transcriptomic level. Regulation networks evolved under different rearrangement rates show huge structural differences, ranging from very small hardly connected networks (high rates) to large and densely connected ones (low rates). Moreover, like in prokaryotes, the number of transcription factors scales quadratically with the number of genes (Beslon et al., 2010).

III) Finally, we showed that this indirect pressure induces many side effects. In particular, under high rearrangement rates, genome compaction causes a fusion of transcribed sequences favouring operons (Parsons et al., 2010).

Thus, by changing a single parameter in the simulations – the spontaneous rearrangement rate – we were able to reproduce genomic and transcriptomic structures ranging from virus-like structures to prokaryote-like and, ultimately, eukaryote-like ones. Moreover, we were able to show that the best final organisms obtained in all these simulations share the same variability level. If we measure the probability for the best final organisms to reproduce neutrally (i.e. the product of its offspring number W by its fraction of neutral offspring F_ν), we always observe that $F_\nu W \sim 1$, showing that these very different organisms all share a same exploration-exploitation compromise, an evident hallmark of indirect selection.

Of course, in Aevol many selective and non-selective effects have been neglected (energetic costs, mutational biases...) that may interact with the indirect selective pressure we isolated. However, in the model, this indirect selective pressure appears to be strong enough to overcome direct selective pressure (high rearrangement rates forbidding organisms to increase their gene repertoire). Thus it is likely to have an effect in real organisms. We now use Aevol to better understand the traces that indirect selection may leave in genomes. We will then be able to search for these traces in the sequences that accumulate in databases.

References

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

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

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.

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

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

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

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

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.