

# Self-Adaptation of Genome Size in Artificial Organisms

C. Knibbe<sup>1</sup>, G. Beslon<sup>1</sup>, V. Lefort<sup>1</sup>, F. Chaudier<sup>2</sup>, and J.-M. Fayard<sup>3</sup>

<sup>1</sup> Prisma lab., INSA Lyon, 69621 Villeurbanne Cedex, France  
guillaume.beslon@insa-lyon.fr,

<sup>2</sup> Biosciences Department, INSA Lyon, 69621 Villeurbanne Cedex, France

<sup>3</sup> BF2I - UMR 0203 INRA/INSA Lyon, 69621 Villeurbanne Cedex, France

**Abstract.** In this paper we investigate the evolutionary pressures influencing genome size in artificial organisms. These were designed with three organisation levels (genome, proteome, phenotype) and are submitted to local mutations as well as rearrangements of the genomic structure. Experiments with various per-locus mutation rates show that the genome size always stabilises, although the fitness computation does not penalise genome length. The equilibrium value is closely dependent on the mutational pressure, resulting in a constant genome-wide mutation rate and a constant average impact of rearrangements. Genome size therefore self-adapts to the variation intensity, reflecting a balance between at least two pressures: evolving more and more complex functions with more and more genes, and preserving genome robustness by keeping it small.

## 1 Introduction

As Maynard-Smith pointed out in 1982, the evolution of large-scale genomic features is “one of the most difficult, perhaps *the* most difficult, question in evolutionary biology” [1]. Since then, molecular biology provided us with huge data about individual genes. Still, little is known about the forces that shape the global structure of the inheritable information in living systems. Experimental evolution of natural systems like cultivable and fast-replicating bacteria takes years [2]. Artificial organisms, by allowing for rapid experiments and parameter control, can help understanding the basic processes at work in evolving systems.

Although genetic algorithms proved useful to study population-level problems, they cannot capture the genome dynamics. Their genotype-phenotype map, where the contribution of each gene relies on its locus, requires both gene number and gene order to be predefined. This forbids changes in genome length and leads to a frozen gene organisation. Pioneering work aiming at removing these constraints [3, 4] kept a fixed phenotypic structure, with a given number of functions, each of them having to be performed by one gene. They therefore needed an external daemon to choose the expressed genes.

Natural genomes owe their degrees of freedom to a flexible phenotypic structure and to the existence of an intermediate level between the genotype and the

phenotype: the set of proteins, whose interactions ensure the survival and reproduction functions. Therefore, to study the evolution of the genomic structure, we introduced such a level into artificial organisms competing for reproduction. Each of them owns a genome encoding basic functional elements, whose interactions produce the phenotype. Both the genomic and the functional structures are evolvable, by the means of punctual mutations and large-scale rearrangements of the genetic material.

Section 2 presents this platform, called *aevol*, notably detailing how it enables us to test self-adaptation hypothesis. In section 3, we focus on the experiments we carried out to test the influence of the mutational pressure on genome size, and we discuss these results in section 4. We conclude in section 5.

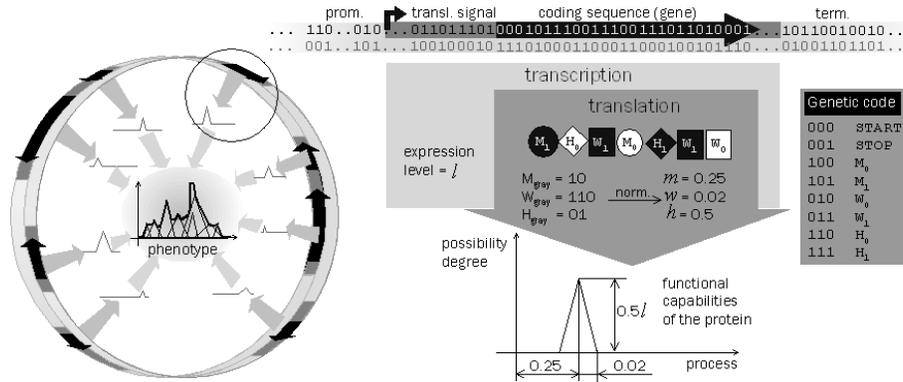
## 2 Designing organisms with flexible genomic and functional structures

The *aevol* system aims at giving as much freedom as possible to let the different levels self-organise. To reach this objective, some features of natural genetic systems were reproduced: (i) the genome is made up of a variable number of genes separated by non-coding sequences, (ii) mutations can modify the genomic structure, (iii) the expression level and the function of a gene are not predefined, and they do not depend on its position but on the local sequence, and (iv) the phenotype results from the interactions of basic functional elements encoded by genes. The remainder of this section describes how these features are implemented.

### 2.1 From genotype to phenotype

As shown by Figure 1, the genome is a circular, double-strand binary string, where 0 and 1 are the complementary bases. Not all the positions are functional: coding sequences (genes) are detected by transcription-translation process inspired by bacterial genetics. Genes are then translated into basic functional elements (proteins). These elementary functions are combined together to get the global abilities of the organism (phenotype).

**Transcription** Sequences called promoters and terminators define both the boundaries and the expression level of the transcribed regions. A preliminary study showed that long *and* frequent terminators associated with rare promoters (i) allow for the emergence of coding sequences, and (ii) limit the overlaps of transcribed regions, thereby giving more freedom for gene rearrangements. Thus, a long consensus sequence was defined to detect promoters (sequences whose Hamming distance with the consensus is  $d \leq d_{max}$ ), whereas terminators are located using their secondary structure ( $abcd^{***}\bar{d}\bar{c}\bar{b}\bar{a}$ ). The expression  $l$  level of a transcribed region depends on the similarity between its promoter and the



**Fig. 1.** Overview of the phenotype computation.

consensus<sup>4</sup>:  $l = 1 - \frac{d}{d_{max} + 1}$ . The following experiments were carried out with a consensus of 28 base pairs (bp) and  $d_{max} = 4$ .

**Translation** During the translation phase, transcribed regions are searched for coding sequences: once a start signal is found, the subsequent positions are read three by three (codon by codon) until a stop signal is reached. The start signal is made up of a Shine-Dalgarno-like sequence followed by the START codon (011011\*\*\*000), and the stop signal is simply the STOP codon (001). Indeed, a longer start signal limits the overlaps between coding sequences, and hence the rigidity of the gene organisation. Once the coding sequences are located, an artificial genetic code is used to translate them into proteins, able to either activate or inhibit processes.

These functional capabilities are expressed within a fuzzy logic framework: the set of processes that can be achieved in our artificial world is an interval of  $\mathbb{R}$  ( $[0, 1]$  here), and each protein is represented by the fuzzy subset of processes it is involved in. This fuzzy formalism enables us to assign a non-null *possibility degree* to each process the protein inhibits or contributes to. The action of a protein can therefore be described by its bell-shaped possibility distribution, approximated by piecewise linear distributions (Figure 1).

Three parameters are necessary to describe such a distribution: its mean  $m$ , its width  $w$ , and its maximal height  $H$ . The main process  $m$  and the interaction potential  $w$  are supposed to depend on the coding sequence only. But the maximal possibility degree  $H$  is limited both by the intrinsic efficiency  $h$  of the protein and by the expression level  $l$  of the region. The genetic code enables us to assign the contribution of each codon to the value of  $m$ ,  $w$  or  $h$ , via a Gray

<sup>4</sup> This simplistic notion of protein quantity was not introduced to model the complex regulation processes at work in living organisms, but rather to allow new gene copies to reduce temporarily their phenotypic contribution, thereby allowing their sequences to diverge.

encoding (Figure 1). The sign of  $h$  determines the activator or inhibitor nature of the protein, and its absolute value is used to compute  $H = l|h|$ .

**Functional interactions** A given process may be achieved by several proteins and inhibited by several others. Therefore, the fuzzy set of processes the organism is able to perform contains the processes that are activated AND NOT inhibited by its proteins. If  $A_i$  is the set of the  $i$ -th activator protein and  $I_j$  the set of the  $j$ -th inhibitor protein, the set of the organism is  $P = (\cup_i A_i) \cap (\overline{\cup_j I_j})$ , and its phenotype is the possibility distribution of  $P$ . Lukasiewicz fuzzy operators are used to perform this combination.

## 2.2 Selection

The environment is also represented by a fuzzy subset of processes, whose possibility distribution is arbitrarily defined. An organism is well adapted if it performs the processes that are feasible in the environment. The higher the gap  $g = \int_0^1 |E(x) - P(x)| dx$  between the environmental distribution  $E$  and the phenotype  $P$  of an organism, the lower its offspring size. The environmental distribution  $E$  we used for the following experiments is shown by Figure 2(b).

The population management is similar to the classical methods used in genetic algorithms. The population size is fixed, and at each time step, all parents are replaced by the offspring. The fitnesses are assigned by linear ranking of  $g$ . The actual selection is done by stochastic sampling with replacement.

## 2.3 Variation operators

The genome of each selected organism is replicated with eventual random errors, affecting a few positions (local mutations) or huge genomic segments (rearrangements), regardless of their function. Genetic exchange between organisms (crossover) is also implemented, but was not used in the following experiments.

Three types of local mutations can be performed at a given position: “switching” its value, inserting or deleting one to six bp. For each mutation type, the number of events per replication follows the binomial law  $\mathcal{B}(L, \mu)$ , where  $L$  is the genome length and  $\mu$  the per-locus mutation rate.

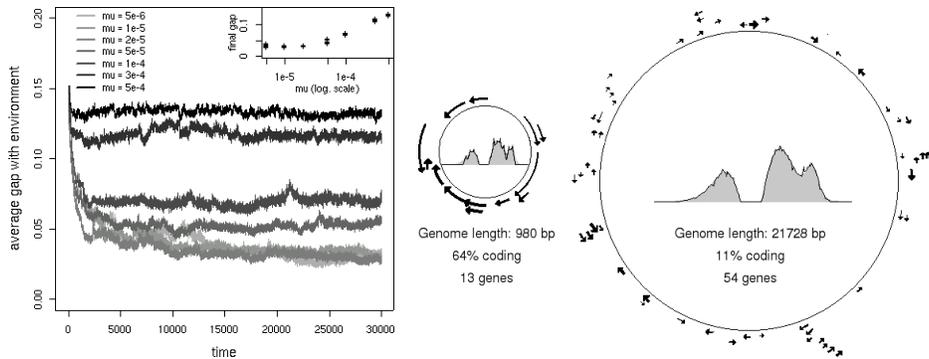
Large-scale rearrangements involve the choice of a genomic segment to be deleted, duplicated, translocated (moved) or inverted. The numbers of events per replication also follow binomial laws. The boundaries of the segment and the eventual insertion point are chosen with a uniform law, edge effects being avoided by the circularity of the genome.

## 2.4 Properties of the system

The proteome level we introduced removes the rigidity of the functional structure: a given process may be achieved by a variable number of functional elements. This in turn removes the rigidity of the genomic organisation; the genome

can undergo rearrangements and indels without preventing phenotype computation. Genome length, gene number and gene order are therefore free to change.

This property enables us to test evolutionary hypothesis involving genome self-organisation through and according to the selection and the variation mechanisms. Indeed, organisms are selected on the basis on their phenotype only, independently of their genomic features: different genome lengths or different gene orders can give the same phenotype, and hence the same fitness. Nevertheless, while the organisms adapt to their environment (Figure 2(a)), some reproducible genomic and functional structures emerge in the long term, for each parameter set. Figure 2(b) shows two different genomic structures, obtained with the same selection method but with different mutation rates.



(a) Evolution of the average gap  $\bar{g}$  between the phenotypes and the environmental distribution, for various per-locus mutation rates ( $\mu$ ).

(b) Best organism after 30000 generations, obtained with  $\mu = 10^{-4}$  (left) or  $\mu = 5.10^{-6}$  (right) for each mutation type. The arrows represent the genes, the black curve the phenotype, and the grey area the environmental distribution  $E$ .

**Fig. 2.** While organisms adapt to their environment, a specific genomic organisation emerges, depending on the mutation rates.

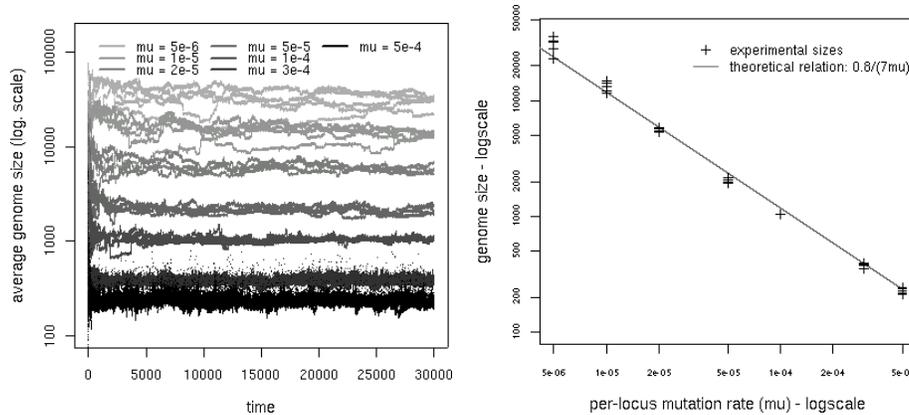
### 3 Experimenting the influence of the mutational pressure on genome size

We used this system to investigate specifically the evolutive pressures acting on genome size. In the experiments we carried out, the seven possible mutations had the same per-locus rate  $\mu$ , ranging from  $5.10^{-6}$  to  $5.10^{-4}$  mutations per bp. For each mutation rate, we let five asexual populations of 1000 artificial organisms evolve independently during 30000 generations, within the steady environment shown in Figure 2(b). The populations were seeded with random genomes of 5000 bp.

### 3.1 Stability of genome size, convergence on local optima

The initial genomes generally do not contain any gene, but after a few generations, local mutations allow a first gene to be expressed. If the set of processes it achieves are feasible in the environment, this first gene is maintained and quickly duplicated. The actions of all copies are summed up, and the organism's abilities eventually exceed the environment's, which is deleterious in our model. Some of the copies are subsequently lost, while other copies diverge: local changes in their coding sequences modify the average process  $m$  they are involved in, allowing the corresponding proteins to move along the functional axis and achieve new processes.

To close the gap with the environmental distribution, the organisms could then adapt the efficiency or the expression level of the genes they already own. Yet a finer tuning could be achieved by acquiring more and more balancing inhibitor/activator genes. However, Figure 3 shows that after a short phase of massive gene acquisition, the genome size reaches an equilibrium, and so the fitness does, trapping the populations on local optima (Figure 2(a)).



**Fig. 3.** Genome size stabilises at a value that depends on the mutation rate  $\mu$ .

When the phenotype is close to the environmental distribution, duplicating a gene becomes more deleterious, which undoubtedly slows down the gene acquisition. Indeed, after the first phase of massive gene acquisition by duplications, the fixation rates of both duplications and large deletions drop to 0.01-0.02 events per generation, whereas other mutation types all stabilise at a higher value, up to 0.15 events per generation, depending on the mutational pressure  $\mu$ . However, it is still theoretically possible to create a gene “from scratch” or to duplicate a coding sequence without its promoter and letting it diverge before expressing it. Therefore, the genome should continue to grow, although more and more slowly.

Moreover, the deleterious effect of the duplications cannot explain simply that the higher the mutation rates, the smaller the equilibrium genome size,

and the lower the average fitness: high mutation rates prevent the organisms from developing complex, highly adapted functions. Therefore, surprisingly, being trapped on local optima is not the consequence of too low a mutation rate. On the contrary, it seems to come from the global pressure of the mutation events on the genome structure, as we shall argue in the following sections.

### 3.2 Convergence towards a constant genome-wide error rate

If the genome size  $L$  varies, then the expected number of mutations per replication  $M = 7\mu L$  changes too. Now, while a low  $M$  prevents the exploration of new solutions, a high  $M$  endangers the robustness of the current one. The existence of a genomic mutation rate, named *error threshold*, “beyond which structures created by an evolutionary process are destroyed more frequently than selection can reproduce them” [5] was demonstrated both in quasi-species models [6] and in genetic algorithms [5]. However, for both models, the genome size is generally fixed, and the mutation rate must be carefully chosen to balance efficiently exploitation and exploration.

Figure 4(a) shows that in our experiments, where genome size is free to vary, the equilibrium size is such that  $M$  takes the same value (around 0.8 mutations per replication), regardless of the mutational pressure  $\mu$ . The equilibrium size can therefore be predicted from  $\mu$  with an hyperbolic relation (Figure 3):  $L \simeq \frac{0.8}{7\mu}$ .

Thus, the genome size stability reflects a compromise between - at least - two contradictory pressures: on the one hand, improving the phenotype with more and more genes, and on the other hand, resisting mutational pressure by keeping genome size small.

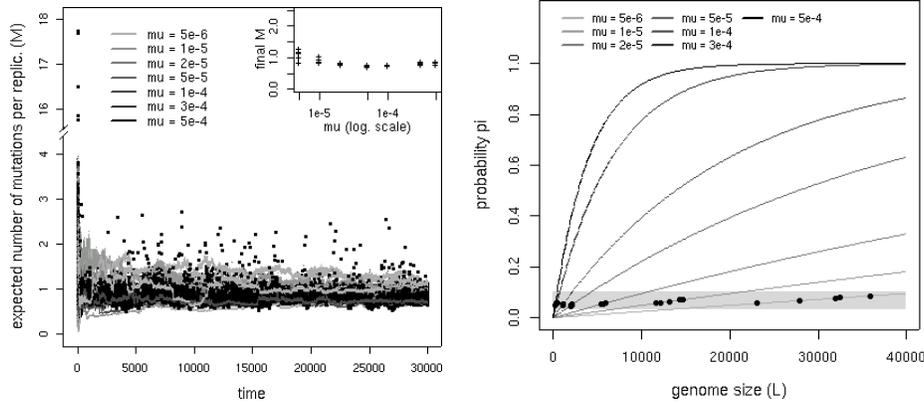
### 3.3 Convergence towards a constant impact of rearrangements

The probability that during a replication, a given position  $\alpha$  is affected by a local mutation is the same ( $\mu$ ) whatever the genome size. On the contrary, the average impact of a rearrangement does increase with genome length  $L$ . Indeed, for a given rearrangement type, say inversion, this impact can be estimated by the probability  $\pi$  for  $\alpha$  to be affected by at least one of the inversions performed during a replication. This probability can be approximated by  $\pi = 1 - \left(1 - \mu \frac{L+1}{2L}\right)^L$  under the hypothesis that the successive rearrangements are independent.

Figure 4(b) shows that in our experiments the genome size spontaneously converges towards the same value of  $\pi$  (around 0.05): the genome size is such that the rearrangement phase of the replication keeps the same average impact when  $\mu$  changes. Besides, figure 4(b) shows that the lower  $\mu$  is, the weaker the slope of  $\pi$  is, which could explain the high run-to-run variability in genome size observed for the low mutation rates (Figure 3).

### 3.4 Changing mutation rates during the evolution

To confirm these results, and to disentangle the effects of small mutations and rearrangements, we carried out additional experiments: we let a population of



(a) The genome-wide error rate  $M$  quickly stabilises at the same value, whatever the per-locus error rate.

(b) Theoretical relation between  $\pi$  and  $L$ . Black points locate the experimental genome sizes.

**Fig. 4.** A need for robustness could explain the limitation of genome size.

artificial organisms evolve during 30000 generations with  $\mu = 10^{-5}$ . Then we changed the rate of the small mutations and/or the rate of the rearrangements, and let the evolution go on.

When all mutation rates are increased, the acquired genomic structure is quickly displaced by a new, shorter one, more robust but less fit. This shows that the shrinkage effect can be surprisingly strong, compared to the pressure for individual adaptation. When, on the contrary, all mutation rates are lowered, this constraint relaxes, enabling the genome to grow. In both cases, final genome sizes can be predicted with the relations  $M = 0.8$  or  $\pi = 0.05$ .

Increasing only the rearrangement rates suffices to make the genome shrink. On the contrary, if we keep high rates for local mutations and reduce rearrangement rates, genome size does not increase. Low rates for both small and large-scale mutations are thus required to make it grow. Analysing the effects of the local mutations, in relation with the coding proportion of the genome, should help us understanding this asymmetric behaviour.

## 4 Discussion

To explain both the diversity and the stability of genome sizes observed in natural organisms, a mutational equilibrium model was recently discussed [7]. This model relies on two different bias: on the one hand, a mutational bias towards small deletions in the indel mechanisms, and on the other hand, a higher fixation rate of large insertions compared to large deletions. Although such bias can exist

in natural species [8], our experiments show that they are not mandatory to stabilise genome size.

Other hypotheses accounting for the steady genome sizes of natural organisms involve natural selection acting directly at the genomic level, either as a stabilising force maintaining the DNA content at a physiological optimum [9], or as a directional pressure counterbalancing the proliferation of “selfish” or “junk” DNA [10, 11] by favouring a short replication time. However, it was shown that there is no correlation between genome size and doubling time among prokaryotes [8]. Besides, in our organisms, genome size does not increase infinitely, although it has no effect on the reproduction rate.

Natural selection can also limit genomic growth a posteriori because of mutational load effects. As mentioned above, the larger a genome is, the more errors occur during its replication. Now quasi-species theory predicts that natural selection favours the set of genotypes, linked by mutation, whose average fitness is highest [12]. It was shown that an evolving population concentrates on the most robust genotypes of the neutral network of high fitness [13]. Experiments with the Avida platform [14] confirmed that digital organisms occupying low but flat fitness peaks can even displace fitter but less robust ones, provided that mutation rates are high enough [15].

In both studies, the genome length was fixed, but what happens if several genotypes with different lengths but the same fitness are in competition? Smaller genomes will undergo less mutations per replication, thus the size of the steps on their peaks will be smaller: everything happens as if they stood on a flatter peak than the larger genomes. Quasi-species theory would therefore predict that under high mutation rates, these smaller genomes will win the competition. This is indeed what our experiments tend to show.

Thus, under high mutation rates, the long-term selection for robustness is clear, quite unlike the genomic growth under low mutation rates. In our system, there is undoubtedly a pressure to evolve more complex functions involving more genes. But longer genomes also undergo more mutations per replication, which can compensate for a low mutation rate and allow for the exploration of new parts of the fitness landscape. A long-term selection for evolvability could therefore also occur.

## 5 Conclusion and future work

Experiments with our artificial system confirmed the intuitive idea that genomic growth, leading to more and more complex phenotypes, can be efficiently limited by a need for robustness. Rearrangements, and especially duplications and deletions, seem to play a key role in this equilibrium: they are the agents of genome length variation, and at the same time, genome length seems to be limited by their average impact.

The selective pressures that actually make the genome grow towards this maximum value have to be investigated further, notably to test the existence of a selection for evolvability, and to understand the role of local mutations in

genomic growth. A detailed study of the effects of each mutation type on the fitness, in relation with the gene number and the coding proportion of the genome, should help us understanding this process. This study should also investigate the influence of the selection intensity on the robustness constraint.

Besides, our system also enables us to study the evolution of gene organisation: since the functional structure and the genomic structure co-evolve, we can analyse the putative retro-actions of the functional level on the gene organisation, notably those leading to genetic modularity.

## References

1. Maynard-Smith, J.: Overview - unsolved evolutionary problems. In Dover, G.A., Flavell, R.B., eds.: *Genome evolution*, New York, NY, Academic Press (1982) 375–382
2. Lenski, R.E.: Phenotypic and genomic evolution during a 20,000-generation experiment with the bacterium *Escherichia coli*. *Plant Breeding Reviews* **24** (2004) 225–265
3. Goldberg, D.E., Deb, K., Kargupta, H., Harik, G.: Rapid accurate optimization of difficult problems using fast messy genetic algorithms. In Forrest, S., ed.: *Proceedings of the Fifth International Conference on Genetic Algorithms*, San Mateo, CA, Morgan Kaufmann (1993) 56–64
4. Burke, D.S., De Jong, K.A., Grefenstette, J.J., Ramsey, C.L., Wu, A.S.: Putting more genetics into genetic algorithms. *Evolutionary Computation* **6** (1998) 387–410
5. Ochoa, G., Harvey, I., Buxton, H.: Optimal mutation rates and selection pressure in genetic algorithms. In: *Proceedings of Genetic and Evolutionary Computation Conference (GECCO-2000)*, San Francisco, CA, Morgan Kaufmann (2000)
6. Eigen, M., Schuster, P.: *The hypercycle: A principle of natural self-organization*. Springer-Verlag (1979)
7. Petrov, D.A.: Mutational equilibrium model of genome size evolution. *Theoretical Population Biology* **61** (2002) 533–546
8. Mira, A., Ochman, H., Moran, N.A.: Deletional bias and the evolution of bacterial genomes. *Trends in Genetics* **17** (2001) 589–596
9. Gregory, T.R.: Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. *Biological Reviews of the Cambridge Philosophical Society* **76** (2001) 65–101
10. Doolittle, W.F., Sapienza, C.: Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284** (1980) 601–603
11. Ohno, S.: So much junk DNA in our genome. In Smith, H.H., ed.: *Evolution of Genetic Systems*, New York, Gordon and Breach (1972) 336–370
12. Eigen, M., McCaskill, J., Schuster, P.: The molecular quasi-species. *Adv. Chem. Phys.* **75** (1989) 149–263
13. Van Nimwegen, E., Crutchfield, J.P., Huynen, M.: Neutral evolution of mutational robustness. *Proc. Natl. Acad. Sci. USA* **96** (1999) 9716–9720
14. Ofria, C., Wilke, C.: Avida: A software platform for research in computational evolutionary biology. *Artificial Life* **10** (2004) 191–229
15. Wilke, C.O., Wang, J.L., Ofria, C., Lenski, R.E., Adami, C.: Evolution of digital organisms at high mutation rates leads to the survival of the flattest. *Nature* **412** (2001) 331–333