Available online at www.sciencedirect.com
ScienceDirect
Journal of Theoretical Biology 244 (2007) 621–630
Journal of Theoretical Biology
www.elsevier.com/locate/yjtbi

Evolutionary coupling between the deleteriousness of gene mutations and the amount of non-coding sequences

Carole Knibbe[a,b], Olivier Mazet[c], Fabien Chaudier[d], Jean-Michel Fayard[b], Guillaume Beslon[a,*]

[a]*Computer Science Department, INSA Lyon, Bat. Blaise Pascal, 69621 Villeurbanne Cedex, France*
[b]*Laboratoire de Biologie Fonctionnelle, Insectes et Interactions, UMR INRA/INSA 203, INSA Lyon, Bat. Louis Pasteur, 69621 Villeurbanne Cedex, France*
[c]*Camille Jordan Institute of Mathematics, INSA Lyon, Bat. Leonard de Vinci, 69621 Villeurbanne Cedex, France*
[d]*Biosciences Department, INSA Lyon, Bat. Louis Pasteur, 69621 Villeurbanne Cedex, France*

Received 31 March 2006; received in revised form 20 July 2006; accepted 6 September 2006
Available online 12 September 2006

Abstract

The phenotypic effects of random mutations depend on both the architecture of the genome and the gene–trait relationships. Both levels thus play a key role in the mutational variability of the phenotype, and hence in the long-term evolutionary success of the lineage. Here, by simulating the evolution of organisms with flexible genomes, we show that the need for an appropriate phenotypic variability induces a relationship between the deleteriousness of gene mutations and the quantity of non-coding sequences maintained in the genome. The more deleterious the gene mutations, the shorter the intergenic sequences. Indeed, in a shorter genome, fewer genes are affected by rearrangements (duplications, deletions, inversions, translocations) at each replication, which compensates for the higher impact of each gene mutation. This spontaneous adjustment of genome structure allows the organisms to retain the same average fitness loss per replication, despite the higher impact of single gene mutations. These results show how evolution can generate unexpected couplings between distinct organization levels.

Keywords: Genome evolution; Non-coding DNA; Mutation effect; Variability; Robustness

1. Introduction

Mutations act on the hereditary material, while selection acts on the morphological and physiological traits of the organism. This is why the relationship between the genotype and the phenotype lies at the heart of evolutionary biology. Although the one-gene-one-character view of early Mendelism was already questioned in Morgan's *The theory of the gene* (Morgan, 1926), this simplification was widely used in classical genetics until the 1960s (Morange, 2000). It had to be abandoned when molecular biology redefined genes as sequences of nucleotides encoding proteins. However, the relationship between proteins and characters still remained mysterious (Morange, 2000). The complexity of this relationship was plainly revealed when progress in high-throughput measurements allowed for the discovery of large and interconnected protein networks, including metabolic, regulatory, signalling or protein–protein interaction networks (Barabasi and Oltvai, 2004). In this complex genotype–phenotype map, polygeny and pleiotropy (Wright, 1968) are the rule rather than the exception: most biological traits result from the interactions of many proteins, and reciprocally, most proteins are involved in several processes.

The genotype–phenotype map plays a critical role in the evolutionary fate of a lineage, because it determines the sensitivity of the phenotype to the mutations in the genes. This sensitivity cannot be too high, otherwise the lineage would quickly die off. There is a long-term pressure to limit the phenotypic variation from one generation to another

*Corresponding author. Tel.: +33 4 72 43 84 87; fax: +33 4 72 43 83 14.
E-mail addresses: carole.knibbe@insa-lyon.fr (C. Knibbe), olivier.mazet@insa-lyon.fr (O. Mazet), fabien.chaudier@insa-lyon.fr (F. Chaudier), jean-michel.fayard@insa-lyon.fr (J.-M. Fayard), guillaume.beslon@insa-lyon.fr (G. Beslon).

0022-5193/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2006.09.005

(Schuster and Swetina, 1988; Wagner et al., 1997; Van Nimwegen et al., 1999; Wilke, 2001; Wilke et al., 2001). At the same time, an ideally robust phenotype, that is, completely insensitive to mutations, would lead to an evolutionary dead end. To sometimes discover beneficial innovations, some non-neutral mutations must happen, otherwise the lineage will lose the evolutionary competition (Layzer, 1980; Fontana and Schuster, 1998; Ancel and Fontana, 2000; Ancel Meyers et al., 2005). Therefore, a certain sensitivity of the phenotype to gene mutations is also necessary, at least transiently. Thus the long-term evolutionary success seems to require an intermediary level of variability, which could put an indirect selective pressure on the general features of the genotype–phenotype map (Wagner and Altenberg, 1996).

The genotype–phenotype map is not, however, the sole level responsible for robustness and variability. The effect of the mutations *in the genes*—or, more generally, in any functional sequence—does depend on the gene–trait relationships, but these mutations represent only a fraction of the mutations the genome undergoes. Indeed, many eukaryotic genomes seem to contain a high proportion of non-functional DNA, and mutations occurring in these sequences are unlikely to affect the phenotype. Therefore, architectural properties of the genome like the proportion of functional sequences are the first filter in the transition from mutation to phenotypic change. Consequently, the genomic structure also contributes to the mutational variability of the phenotype. If a specific level of variability is indirectly selected, variations in the gene–trait relationships may thus induce compensatory changes in the genomic structure, and conversely. Both levels cannot be considered independently.

A systemic understanding of evolving organisms must, therefore, take into account not only the interactions between gene products, but also, at a more integrated level, an interplay between the architecture of the genome and the genotype–phenotype map. Investigating this interplay in natural organisms is, however, a real challenge. Indeed, even within a single cell, the functional interactions between proteins are incredibly numerous and complex and the general features of the genotype–phenotype map can hardly be controlled. Thus, we use here simulation to demonstrate how these general features can influence the structure of the genome.

The evolutionary model we used, called *aevol*, was designed to address general issues about genome evolution, by taking into account some of the biological properties that give natural genomes their degrees of freedom. It is thus fairly different from the genetic algorithms commonly used in computer science. In the classical implementation of genetic algorithms, each position in the genotype is a "gene" with a predetermined function, which implies that gene number and gene order are fixed. To study the evolution of genome organization, gene density, gene number and gene order must, on the contrary, be evolvable. Some tricks were proposed to achieve this within the framework of a genetic algorithm (see for instance Goldberg et al., 1993), but we have chosen to use a more natural way. In our system, a gene is a sequence of positions bounded by specific signals (and not a single absolute position), and the phenotypic contribution of a gene depends on its sequence (and not on its locus). Gene number, gene density and gene order are thus free to evolve by the means of both local mutations and large-scale genomic rearrangements.

To study the possible coupling between such a flexible genome and the general features of the genotype–phenotype map, we also need a mapping with specific properties, in which the genes would contribute to different processes (but not all) and each process is achieved by interactions between different genes (but not all). In other words, we need a mapping that allows for pleiotropy and polygeny, but not universal ones. In real organisms, the proteome and its complex interactions fulfill this role. However, it is clearly impossible to mimic all biochemical interactions in a computational model. In our model, we use an abstract description of "protein" function which obviously prevent direct comparison with real organisms but provides our artificial "organisms" the minimal requirements for our questioning.

In the model, it is possible to control the general features of the genotype–phenotype map by setting the maximal gene pleiotropy. It allows us to modulate the degeneracy of the genotype–phenotype map: the higher the pleiotropy, the more polyvalent the proteins and the more degenerate the mapping. By allowing several populations to evolve under various pleiotropy levels, we observed that the evolved genomic structure depended on the degeneracy of the mapping. Under high pleiotropy, when most gene mutations are highly deleterious, the evolved genome surprisingly contained a lower amount of non-functional sequences. This lowered the average number of genes affected by each rearrangement and restored the average fitness variation per replication. This is, therefore, an example of a long-term interaction between the structure of the genotype–phenotype map and the architecture of the genome.

2. The *aevol* model

2.1. General principles

The basic assumptions of the *aevol* model are that (i) the reproductive success of an organism depends on the adequacy between its phenotype and the environment, (ii) this phenotype can be described by the organism's ability to perform processes, (iii) each performable process is the result of the interactions of functional elements—"proteins"—encoded by genes and (iv) mutational events can affect the number of genes, their positions on the chromosome, their expression levels, the set of processes they are involved in, the strengths of their interactions as well as the lengths of the intergenic regions. An overview of the transition from the genetic material to the phenotype is

shown in Fig. 1 and will be detailed in the following paragraphs. The general idea is that coding sequences are detected using transcription and translation signals and that the translation process defines the functional capabilities of each gene product, as a more or less wide range of (abstract) processes. The combination of all gene products determines the global functional capabilities of the organism, which we call here the phenotype.

The aim of this model is to investigate the minimal conditions that allow for complex evolutionary relationships to emerge between the genotype–phenotype map and genome structure. Modelling precisely a particular gene network or specific biochemical reactions is, therefore, not the purpose. Here, the functions of a protein are not described by precise chemical reactions, but rather by a set of abstract "processes" it can contribute to or inhibit, using the framework of fuzzy set theory (Dubois and Prade, 1980). The global set of processes that can be achieved is an interval of $\mathbb{R}$ ([0, 1] here). Each protein can either contribute to or inhibit a fuzzy subset of these processes. Protein can then interact in a general, functional sense, if they are involved in common processes, that is, if the intersection of their fuzzy sets of processes is not empty. This formalism does not pretend to represent accurately the genotype–phenotype map of any real organisms—for instance, it is rather unlikely that real biological processes can be ordered on a one-dimensional, continuous space. However, this formalism allows for a genotype–phenotype map with both pleiotropic genes and polygenic traits while remaining human-understandable and computationally tractable. It allows us to study the evolutionary pressures that take place in an evolving system that includes a realistic genome structure and a non-trivial genotype–phenotype map, while being kept as simple as possible.

2.2. From the genetic material to the phenotype

2.2.1. Transcription

As shown by Fig. 1, the chromosome is a circular, double-strand, binary string, where 0 and 1 are the complementary bases. In this chromosome, the transcription phase searches for sequences called promoters and terminators to define the boundaries of the transcribed regions. These sequences are inspired from biological signals: sequences similar to a consensus sequence are promoters, while sequences able to form a stem–loop structure ($abcd***\bar{d}\bar{c}\bar{b}\bar{a}$) are terminators.

The promoter sequence also contributes to the expression level of the subsequent coding sequences: the quantity of proteins increases when the promoter resembles the consensus sequence. In our model, this rough modulation of the expression level was introduced to allow duplicated genes to reduce temporarily their phenotypic contribution and let their sequences diverge toward other functions. The number of differences d between the promoter and the consensus sequence is used to compute the expression level $e = 1 - d/(d_{max} + 1)$. All experiments were carried out with a consensus of 28 base pairs (bp) and $d_{max} = 4$.

2.2.2. Translation

The translation phase searches for coding sequences inside transcribed regions: 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). Note that this transcription–translation process allows for operons. Indeed, although we study here only the most basic features of genome structure such as gene number and gene density, we plan to investigate the evolution of operons and gene clusters in relation with the genotype–phenotype map, as the next step of our research. Once a coding sequence is found, an artificial genetic code is used to translate it into a protein, able to either activate or inhibit a fuzzy subset of processes. This fuzzy subset is represented by a piecewise-linear possibility distribution with a "triangular" shape (Fig. 1). Such distributions can be described by three parameters: the mean m, the width w, and the maximal height H. This basically means that the protein is involved in the processes

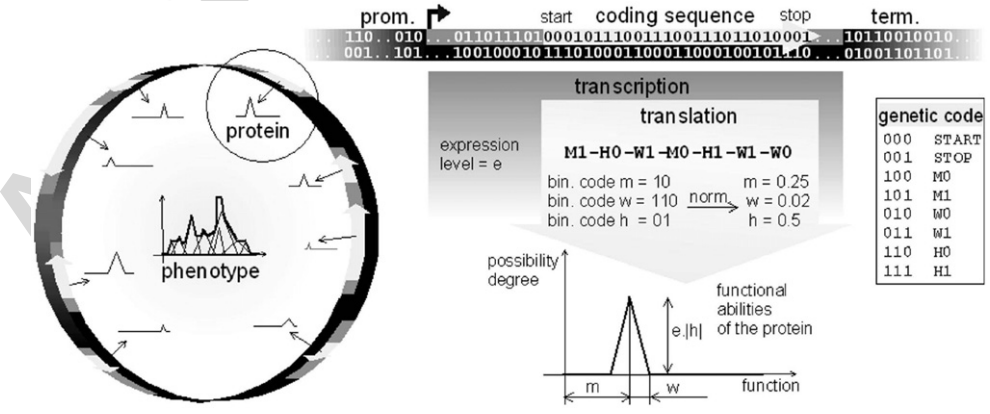

Fig. 1. Overview of the phenotype computation in the *aevol* model. Coding sequences are detected using signal sequences and translated using an artificial genetic code. The translation step uses a genetic code to characterize the subset of processes activated or inhibited by the protein. The phenotype of the organism, describing its global functional capabilities, results from the interactions of all the proteins.

$[m-w, m+w]$ with a varying degree. The mean m defines the main process the protein contributes to. The width w quantifies the pleiotropy of the protein: the higher w, the more processes the protein is involved in. The height H defines the degree with which the protein contributes to the main process. This degree is limited both by the expression level e of the region and by the intrinsic efficiency h of the protein: $H = e|h|$.

The artificial genetic code enables us to compute the values of m, w or h. In a coding sequence, the codons that relate to a given parameter, for instance w, are viewed as a Gray encoding of this parameter (the Gray encoding is a variant of the classical binary encoding). For instance, in the case shown by Fig. 1, three codons contribute to the value of w: W_1, then W_1 again, and finally W_0. The Gray encoding of w is, therefore, 110. This value is then converted to an integer (4 here) and normalized between $[0, w_{max}]$, which finally yields $w = 0.02$ ($w_{max} = 0.033$ in this example). The conversion process is similar for m and h but they are normalized between $[0, 1]$ and $[-1, 1]$, respectively. The sign of h determines whether the protein is activator or inhibitor, and its absolute value is used to compute the maximal possibility degree $H = e|h|$.

The upper bound for w, namely w_{max}, enables us to adjust the maximal gene pleiotropy. A low w_{max} implies that all genes are specialized on a tight range of biological processes, while a higher value allows polyvalent genes to appear (note, however, that specialized genes are still possible). Changing w_{max} should, therefore, modify the average deleteriousness of gene mutations. We used values comprised between 0.01 and 0.33 in the following experiments, to investigate the effects of various pleiotropy levels on genome evolution.

2.2.3. Functional interactions between proteins

We define here a functional interaction between two proteins as an overlap in their sets of processes: two proteins interact if they are involved in common processes. We have thus two ways to represent the proteome. We can either superimpose the possibility distributions of the proteins on the axis of processes, or draw the network of functional interactions between proteins. The latter is a convenient representation because it is commonly used, but it should be kept in mind that the network we obtain cannot be directly compared to a real network of physical protein–protein interactions for instance. We use the term "interaction" in a more general, functional sense. A link in our network could be viewed as the participation to a same metabolic pathway or signalling cascade, for instance.

2.2.4. Phenotype computation

The phenotype of an organism is represented in our system by the fuzzy set of processes it is able to perform. Since a given process may be achieved by several proteins and inhibited by several others, the organism's fuzzy set contains the processes that are activated AND NOT inhibited by its proteins. More formally, if A_i is the set of the ith activating protein and I_j the set of the jth inhibiting protein, the set of the organism is thus $P = (\bigcup A_i) \cap \overline{(\bigcup I_j)}$, and its phenotype is the possibility distribution of P.

Lukasiewicz fuzzy operators are used to perform this combination: the possibility distributions of $\bar{A_1}$, $A_1 \cup A_2$ and $A_1 \cap A_2$ are the functions $x \mapsto 1 - A_1(x)$, $x \mapsto min(A_1(x) + A_2(x), 1)$ and $x \mapsto max(A_1(x) + A_2(x) - 1, 0)$, respectively. This procedure can be intuitively explained as follows. To know what processes the organism can perform and to what degree it can perform them, we sum up the possibility distributions of the activator proteins and we subtract the distributions of the inhibitory proteins. This is actually a little more complicated because a possibility degree cannot be negative or higher than 1. These threshold effects cause some interactions to be non-additive.

2.3. Selection

The fitness of an organism depends on the adequacy between its phenotype and the environment. The environment is also represented by a fuzzy subset of processes E, whose possibility distribution is arbitrarily defined. This fuzzy subset can be seen as the set of processes that are required to survive 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 fewer offspring this organism will have. Fig. 2 shows the graphical interpretation of the gap g, as well and the environmental distribution we used in these experiments.

The population management is similar to the classical methods used in genetic algorithms. The population size, N, is fixed. At each time step, the N organisms are sorted according to the adaptation measure g. Each organism gets

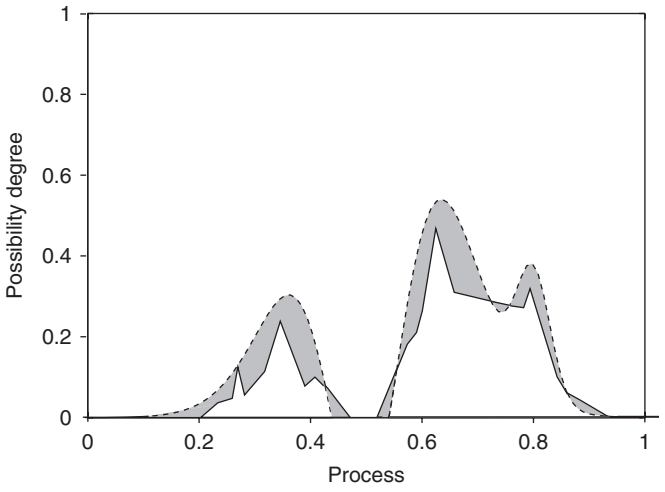

Fig. 2. Graphical interpretation of the adaptation measure g. The dashed line represents the functional capabilities that had to be achieved to survive and reproduce in the environment (possibility distribution E in the text). The continuous line represents an hypothetical phenotype. The "gap" g between both curves, measured by the gray area, is used as a measure of the adequacy between the phenotype and the environment.

a probability of reproduction that is proportional to its rank in the list. Then the number of offspring of each organism is chosen by multinomial sampling. This ranking selection scheme ensures that the selective pressure is constant throughout the evolution (Whitley, 1989). After producing offspring, parents die, so that generations are discrete. Note that since the fitness computation does not include genomic criteria like genome length or gene order, there is no direct selective pressure on genome structure.

2.4. Mutations

Replicating genomes can undergo point mutations, small insertions and small deletions, but also genomic rearrangements in the broad sense, including duplications, deletions, translocations and inversions. At each replication, the number of point mutations (resp., insertions, translocations, ...) is drawn from the binomial law $\mathscr{B}(L,\mu)$, where L is the genome length and μ the per-base rate of point mutation (resp., insertion, translocation, ...). For example, if μ is set to 10^{-5} for each mutation type, a genome of length $L = 10,000$ bp undergoes on average 0.1 point mutation, 0.1 small deletion, 0.1 small insertion, 0.1 duplication, 0.1 large deletion, 0.1 inversion and 0.1 translocation per replication. The boundaries of the rearranged segments (as well as the insertion point for the duplications and the translocations) were chosen independently and uniformly on the chromosome. This uniform law aims at simply modelling the diversity of the molecular mechanisms at work in natural genomes. Indeed, in natural organisms, several recombination mechanisms act at different distances (Rocha, 2003), resulting in an unknown but probably complex size distribution of the rearrangements.

The following experiments were carried out with $\mu = 10^{-5}$ for each mutation type. Genetic exchange between organisms was not used here.

3. Results

To investigate whether genome evolution is influenced by the average deleteriousness of gene mutations, we carried out experiments with six values of the maximal gene pleiotropy w_{max}: 0.01, 0.02, 0.033, 0.1, 0.2 and 0.33. For each value of w_{max}, we allowed five asexual populations of 1000 organisms to evolve independently during 40,000 generations, within the steady environment shown by Fig. 2. Each population was seeded with 1000 random genomes of 5000 bp.

3.1. Single gene mutations are more deleterious when a higher pleiotropy is allowed

Since the initial random genomes generally do not contain any gene, the protein set has to be built up "from scratch". After a few generations, a first functional gene appears. If it is beneficial, this first gene is maintained and quickly duplicated, generally with neighboring non-coding sequences. This results in a sudden increase in genome size. Then local changes in the copied sequences modify the processes they are involved in, allowing the corresponding proteins to move along the functional axis and achieve new processes. Other copies as well as some non-coding parts are deleted and the genome size stabilizes. The initial phase of the evolution always follows this succession of duplication–divergence events, whatever w_{max}.

However, when the maximal gene pleiotropy is changed, the organisms do evolve different types of proteomes. As shown by Fig. 3, when w_{max} is low, most proteins are activator and specialized on a narrow range of processes. When increasing w_{max}, the proteome gradually shifts toward a complex interaction network of activatory and inhibitory proteins, most of them being involved in a wide range of processes. Gene mutations should, therefore, be more deleterious when w_{max} is high. To test this hypothesis, we performed systematic single gene knock-outs on the final fittest organism of each run and evaluated the resulting variation in the adaptation measure g. As shown by Figs. 3 and 4a, the average effect of gene mutations does indeed increase with the maximal gene pleiotropy. This is not a surprising result—this is actually the expected behavior of the system. It was, however, necessary to check that, by acting on w_{max}, we actually modified the average deleteriousness of gene mutations, since this is the factor that has a biological sense.

3.2. The global mutational variability of the phenotype is maintained

If, as shown by Fig. 4a, the sensitivity of the phenotype to gene mutations is increased when highly pleiotropic genes are allowed, one would expect the evolutionary fate of the lineage to be compromised. This would create an indirect selective pressure against pleiotropy, thereby favoring modularity (Wagner and Altenberg, 1996). However, the effect of single gene mutations is only a component of the global fitness variation per replication, which is what selection actually sees. Indeed, the global fitness variation per replication also depends on the mutation rates and on the number of genes affected by each mutation. Remember that mutations can occur inside intergenic regions and be neutral, or, on the contrary, affect several genes at once (when the mutation is a large deletion for instance). Therefore, if for some reason pleiotropy can only hardly be changed, there are in principle other ways to bring back the global variability in the acceptable range.

Here, to estimate the evolved fitness variation per replication, we simulated 50,000 replications of each final fittest organism and compared the offspring to their progenitor. This analysis revealed that despite the higher impact of single gene mutations (Fig. 4a), the expected fitness loss per replication does not increase with w_{max} (Fig. 4b). This is not an artifact of our ranking

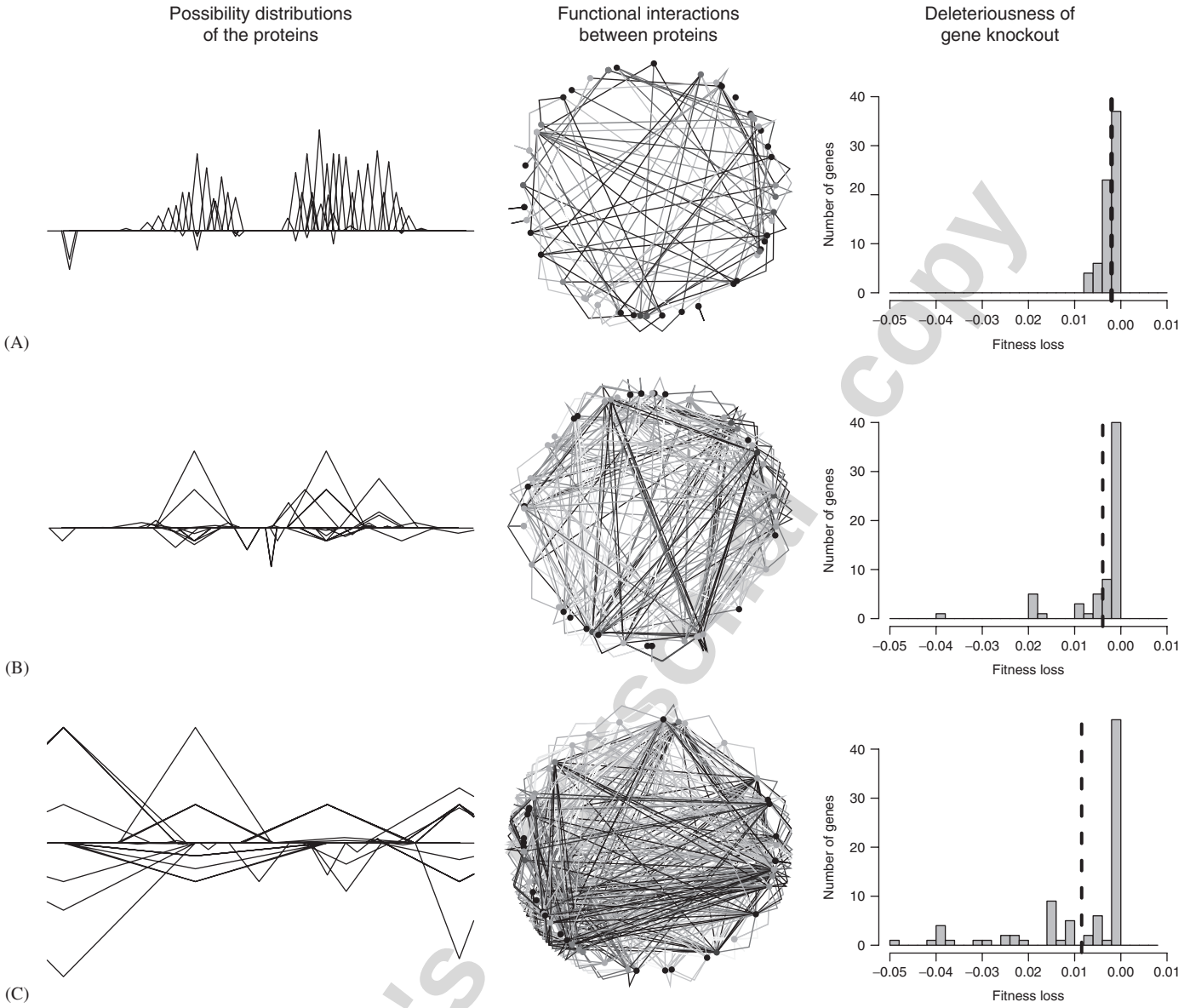

Fig. 3. Examples of evolved "proteomes". The figure shows the analysis of the protein set of the final fittest organism in a representative run, for different values of the maximal gene pleiotropy (A: $w_{max} = 0.02$, B: $w_{max} = 0.1$, C: $w_{max} = 0.3$). Left: superimposition of the possibility distribution of the proteins. "Triangles" with positive (resp., negative) heights represents activator (resp., inhibitory) proteins. When allowing for a higher gene pleiotropy (w_{max}), the proteome shifts from a sum of specialized proteins to a combination of activator and inhibitory proteins, most of them being involved in a wide range of processes. Middle: nodes represent proteins and are located according to the position of the coding sequence on the chromosome. Edges represent functional interactions between proteins (overlapping "triangles"). In our model, increasing gene pleiotropy causes the network of functional interactions between proteins to be more densely connected. Right: histogram of the fitness variation caused by systematic single gene knock-outs. Each gene in the final fittest organism was silenced and the variation in the adaptation measure g was measured. The dashed vertical line in the resulting histogram indicates the mean variation. This experiment confirms that the deleteriousness of gene mutations can be controlled by the parameter w_{max}.

selection scheme, because it also holds if the populations evolve under a different selection scheme, where the reproduction probabilities directly depend on the absolute value of the adaptation measure g (data not shown). This suggests that a specific mutational variability of the phenotype has been maintained in some lineages despite the higher deleteriousness of gene mutations and that these lineages were the successful ones in the evolutionary competition.

To understand this result, we took advantage of the exact knowledge of the mutational events provided by the simulation platform. For each final fittest organism, we analysed the 50,000 simulated replications and counted the number of genes modified by at least one mutational event. As shown by Fig. 5, the higher the maximal gene pleiotropy, the fewer genes were affected. Therefore, when a higher pleiotropy is allowed, mutations in the genes are more deleterious but fewer genes are mutated. This compensatory effect explains that when the whole replication process is taken into account, the average fitness loss does not depend on the maximal gene pleiotropy (Fig. 4b). It is worth noting that the per-base mutation rate used to

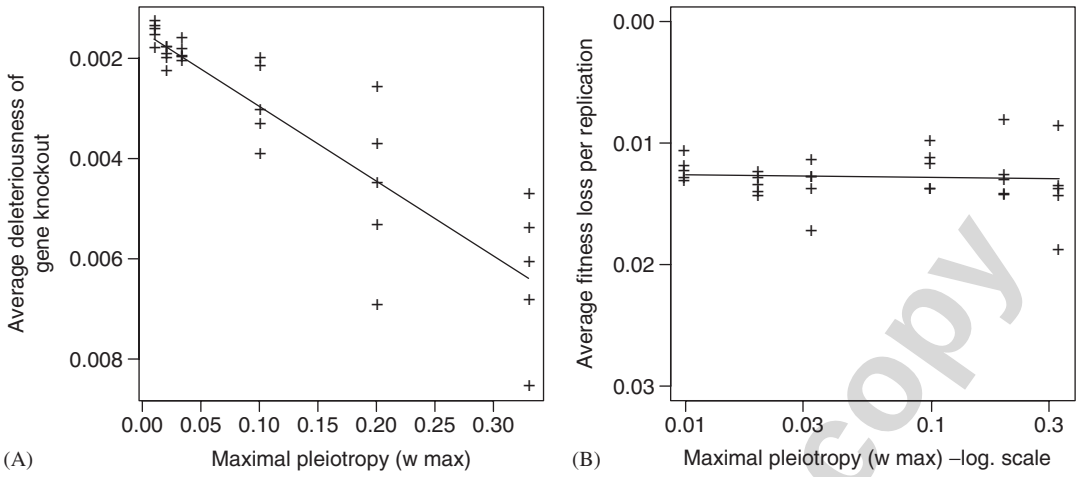

Fig. 4. The global mutational variability of the phenotype is maintained despite the higher impact of gene mutations. (A) Complete results of the knock-out experiments (see Fig. 3). Allowing for a higher gene pleiotropy significantly increases the average deleteriousness of gene mutations ($r^2 = 0.79$, $p = 4 \times 10^{-11}$). (B) However, surprisingly, the average fitness variation per replication remains the same ($r^2 = 0.003$, $p = 0.769$). This fitness variation was obtained by simulating 50,000 independent replications of each final fittest organism with the same mutation rates and mutation operators as during the main run. Note that in each case, the average variation is negative, because most mutations are here either neutral or deleterious, as in natural organisms (Kimura, 1983).

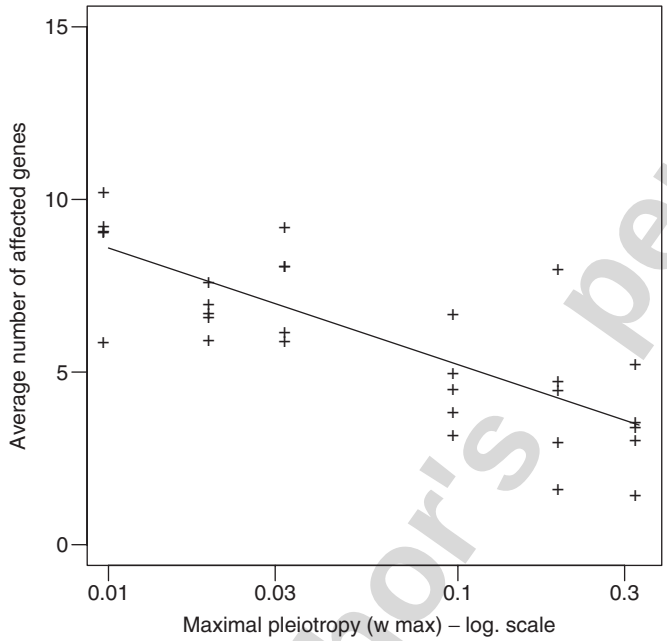

Fig. 5. When gene mutations are more deleterious, fewer genes are mutated per replication. The average number of mutated genes per replication was obtained by analysing the 50,000 simulated replications of the final organisms (see Fig. 4). It is significantly lower when a higher gene pleiotropy is allowed ($r^2 = 0.61$, $p = 3 \times 10^{-7}$). This compensates for the higher fitness loss caused by each mutation.

simulate the replications was the same for all organisms, regardless of the pleiotropy level. Thus, the relationship between the maximal gene pleiotropy and the number of mutated genes is not due to variations is the mutational pressure. As we shall argue below, it rather comes from variations in the genomic structure.

3.3. Changes in the genome structure compensated for the higher deleteriousness of gene mutations

Examination of the evolution of genome architecture on the line of descent of each final fittest organism reveals that the total number of genes carried by the chromosome does not increase indefinitely (Fig. 6a). An equilibrium is reached after a varying time, depending on the maximal gene pleiotropy allowed. As a result, the final gene number is the lowest for the highest pleiotropy levels (Fig. 6b).

One might argue that this relationship is no more than the logical result of allowing for larger triangles: when w_{max} is increased, fewer proteins are required to fill the total area of the environmental distribution with equal precision. However, since the environmental distribution we chose is not piecewise-linear, it cannot be approximated precisely by a few large triangles. An infinite number of triangles is actually required for a perfect fit. Moreover, because of our selection scheme, the selective pressure is maintained even when the population is close to the optimum. Therefore, in these experiments, there is a constant need for additional inhibitory or activator triangles to refine the organism's phenotype. If selection for fitness was the sole force driving the evolution of genome structure, the gene number should not stabilize, whatever the maximal gene pleiotropy. The reduction in gene number is actually one aspect of a more general process of genome shrinkage. Indeed, as a higher gene pleiotropy is allowed, the quantity of non-coding bases at equilibrium also decreases (Fig. 7).

Such an evolutionary coupling is surprising because in our model, non-coding sequences do not contribute to the protein network, the phenotype and the fitness of the organism. However, they do contribute to the variability of the phenotype. Indeed, in a genome with longer intergenic sequences, the probability to find pairs of similar sequences

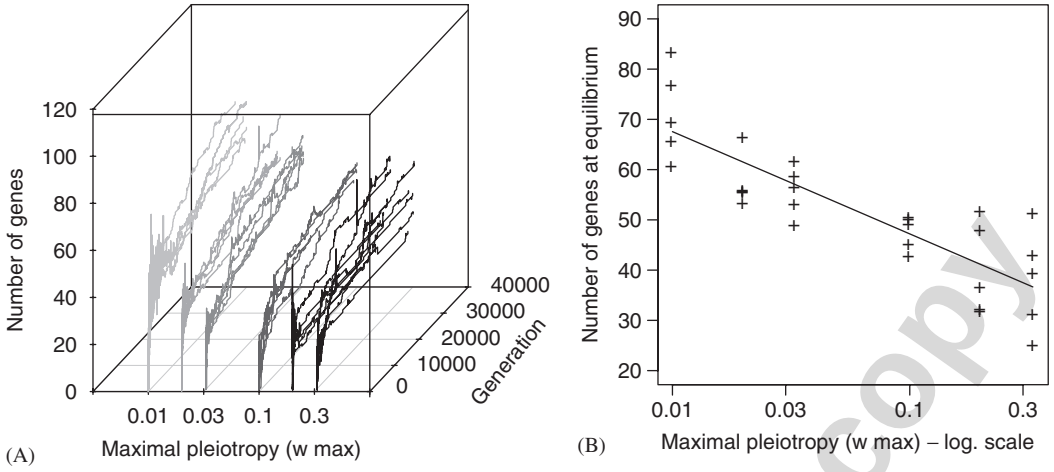

Fig. 6. The evolved genome contains fewer genes if gene mutations are more deleterious. (A) Evolution of gene number on the line of descent of the final fittest organism. After a varying time, the gene number reaches an equilibrium. (B) The gene number at equilibrium decreases with the maximal gene pleiotropy ($r^2 = 0.71$, $p = 6 \times 10^{-9}$). The equilibrium values were estimated by the mean of the last 10,000 generations.

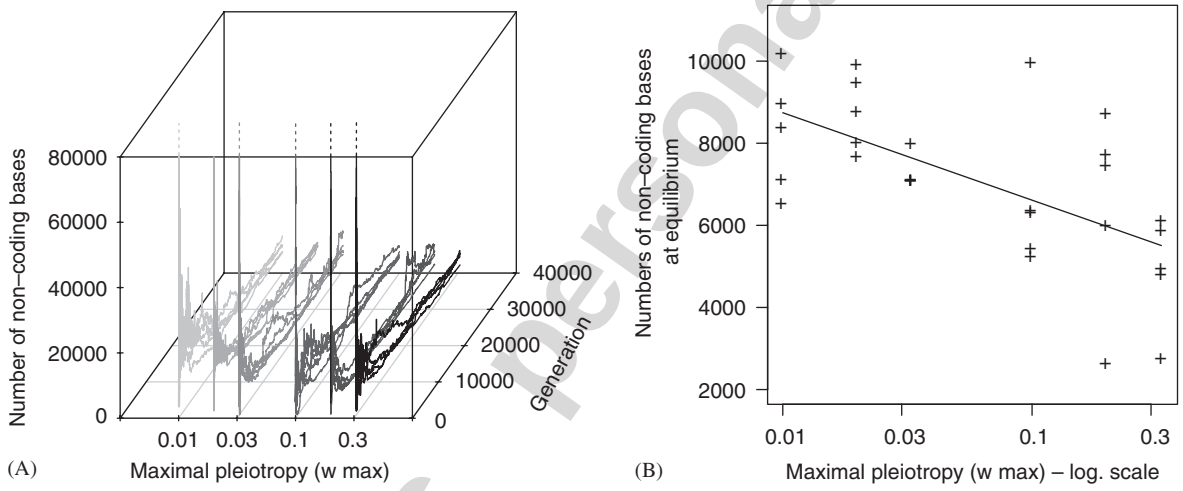

Fig. 7. The amount of non-coding sequences also depends on the deleteriousness of gene mutations. (A) Evolution of the number of non-coding positions on the line of descent of the final fittest organism. After an initial burst, the amount of non-coding sequences also reaches an equilibrium. (B) The amount of non-coding sequences at equilibrium significantly decreases with the maximal gene pleiotropy ($r^2 = 0.38$, $p = 3 \times 10^{-4}$). The equilibrium values were estimated by the mean of the last 10,000 generations.

increases, and hence more genomic rearrangements can occur. Here, this was taken into account in a simple way, by performing, on average, a number of rearrangements proportional to genome length (Section 2.4).

For rearrangements like inversions or translocations, the deleterious effects are concentrated on two or three breakpoints, regardless of the genome compactness. As a consequence, the higher number of events caused by longer intergenic sequences is compensated by an increased probability that the event is neutral. For instance, the probability that an inversion does not affect any gene is $v_{inv} = (1 - l/L)^2$, where L is the genome size and l the number of genic bases. If intergenic sequences grow, the number of inversions per replication increases but they have more chances to be neutral, because l/L drops.

On the contrary, for duplications and deletions, there is no such compensation. All genes contained in the deleted or duplicated segment are affected and the average size of the segment increases when the genome grows. As a result, longer intergenic sequences do not improve the probability that a duplication or a deletion is neutral. Indeed, the probability that a deletion does not affect any gene is $v_{del} = (1/2L^2)\sum_{i=1}^{N_G} \lambda_i(\lambda_i + 1)$, where N_G is the gene number and λ_i the length of the intergenic sequence between genes i and $i+1$. This probability does not tend to 1 when the intergenic sequences grow. For example, it plateaus at $1/2N_G$ if genes are regularly distributed on the chromosome, or at $\frac{1}{2}$ if they form a single cluster.

To sum up, the phenotype is more variable if the genome is longer because duplications and deletions are more numerous and affect on average more genes. In other

words, the average number of duplicated or deleted genes per replication can be increased or decreased by changing the length of the intergenic sequences. Such changes in the genome structure enable the organisms to maintain an appropriate level of phenotypic variation despite the higher deleteriousness of gene mutations, as shown in Fig. 4.

4. Discussion

By revealing an evolutionary coupling between the deleteriousness of gene mutations and the quantity of non-coding sequences, this study shows that changes at a given level can induce unexpected effects at other levels. These effects stem from the common involvement of both the genome architecture and the genotype–phenotype map in the mutational variability of the phenotype. Because the long-term evolutionary success of a lineage requires both a sufficient robustness and a sufficient variability of the phenotype, an intermediate level of phenotypic variation must be maintained. By allowing for higher gene pleiotropy, we caused gene mutations to be more deleterious and hence we perturbed this balance. Compensatory changes in the genome structure restored the appropriate phenotypic variability.

This does not imply that evolution is farsighted. As it was previously demonstrated with other models, lineages where the phenotype is not robust enough simply go to extinction (Schuster and Swetina, 1988; Van Nimwegen et al., 1999; Wilke, 2001; Wilke et al., 2001), while lineages where the phenotype is not variable are outcompeted by those able to generate innovations (Fontana and Schuster, 1998; Ancel and Fontana, 2000; Ancel Meyers et al., 2005). This indirect selection of an appropriate variability of the phenotype is not new in that respect. However, in previous studies, the effect of this selective pressure on genome structure could not be seen because genome length was fixed. Here, by giving more degrees of freedom to the genome, we show that the lineages that survive in the long term are those where the genome structure suits the deleteriousness of gene mutations, ensuring that some offspring retain the ancestral phenotype while others explore new possibilities. Thus, for the organisms we observe after thousands of generations of evolution, the amount of non-coding sequences depends on the deleteriousness of gene mutations, but again, this is not the result of an anticipatory process.

A necessary condition is, however, that the non-coding sequences play a role in the phenotypic variability. This requires that (i) longer genomes undergo more rearrangements and (ii) the average length of a rearranged segment increases with genome size. For both prokaryotes and eukaryotes, the density of repeats is positively correlated with chromosome size (Achaz et al., 2001, 2002), which suggests that longer chromosomes will indeed undergo more recombination events. Besides, since the frequency of homologous recombination does not seem to depend on the distance between repeated sequences (Hughes, 1999), the average length of the rearranged segments is expected to increase with chromosome size. This suggests that in real species genome compactness has indeed an effect on the mutational variability of the phenotype, which is a prerequisite for the coupling with the deleteriousness of gene mutations.

Could such a coupling be revealed by comparing different living species? It is easy now to get information about genome structure, at least for model species. Characterizing the genotype–phenotype maps or the average deleteriousness of gene mutations is a much harder task. Mutation–accumulation experiments have been performed on a number of model species. They suggest that contrary to what we have observed, mutations are more deleterious in the species with the highest gene number (Martin and Lenormand, 2006). However, the direct comparison of different species can be misleading, because distinct traits are used to estimate the fitness (Bataillon, 2000; Martin and Lenormand, 2006). It is for instance the intrinsic growth rate for *Escherichia coli*, the lifetime reproductive success for *Arabidopsis thaliana* and the egg-to-adult viability for *Drosophila melanogaster*. Another fundamental problem with the mutation–accumulation experiments is that they can fail to detect small fitness variations (Bataillon, 2000) as well as huge variations that are counter-selected in microorganisms (Kibota and Lynch, 1996; Martin and Lenormand, 2006).

An indirect approach would consist in comparing the networks of functional interactions between genes. However, even for the most studied species, we only know some pieces of this complex network—for instance, the physical interactions between proteins, the metabolic pathways, the regulatory interactions—but the complete picture is still missing. Besides, even if the complete networks of functional interactions were known for some real species, we do not think that it would be sufficient to predict the effect of gene mutations, because real interactions are too complex. Although there seems to be a relationship between node connectivity and deleteriousness of gene mutations in real networks, it is not as straightforward as it is in our simulations (Jeong et al., 2001; Wuchty, 2004). It seems, therefore, very difficult to compare the deleteriousness of gene mutations across species.

Aside from this difficulty, there is another major reason that would prevent species comparison to reveal the observed coupling. The deleteriousness of gene mutation is far from being the sole factor that distinguishes two real species. Many other factors can differ and have their own effect on genome structure. For instance, we have already shown that genome compactness also critically depends on the per-base mutation rates (Knibbe et al., 2005). In eukaryotes, the proliferation of transposable elements is another additional pressure on genome size. Comparing distant species will almost certainly fail to reveal directly the relationship we have underscore here. This is precisely where the interest of such a modelling approach lies. It enabled us to isolate one pressure on genome structure that

is tangled with many others in living species. Indeed, this approach enabled us to show that all else being equal, the indirect selection of a specific mutational variability leads to a coupling between the deleteriousness of gene mutations and the amount of non-functional DNA.

Acknowledgment

The authors thank A. Coulon for comments on the manuscript. This work is part of the BSMC project (Systems Biology and Cell Modelling) and is supported by the Rhone-Alpes region.

References

Achaz, G., Netter, P., Coissac, E., 2001. Study of intrachromosomal duplications among the eukaryote genomes. Mol. Biol. Evol. 18, 2280–2288.

Achaz, G., Rocha, E.P., Netter, P., Coissac, E., 2002. Origin and fate of repeats in bacteria. Nucleic Acids Res. 30, 2987–2994.

Ancel, L.W., Fontana, W., 2000. Plasticity, evolvability, and modularity in rna. J. Exp. Zool. 288, 242–283.

Ancel Meyers, L., Ancel, F.D., Lachmann, M., 2005. Evolution of genetic potential. PLoS Comput. Biol. 1, e32.

Barabasi, A.-L., Oltvai, Z.N., 2004. Network biology: understanding the cell's functional organization. Nat. Rev. Genet. 5, 101–113.

Bataillon, T., 2000. Estimation of spontaneous genome-wide mutation rate parameters: whither beneficial mutations? Heredity 84, 497–501.

Dubois, D., Prade, H., 1980. Fuzzy Sets and Systems Theory and Applications. Academic Press, New York.

Fontana, W., Schuster, P., 1998. Continuity in evolution: on the nature of transitions. Science 280, 1451–1455.

Goldberg, D.E., Deb, K., Kargupta, H., Harik, G., 1993. Rapid accurate optimization of difficult problems using fast messy genetic algorithms. In: Forrest, S. (Ed.), Proceedings of the Fifth International Conference on Genetic Algorithms. Morgan Kaufmann, San Mateo, CA, pp. 56–64.

Hughes, D., 1999. Impact of homologous recombination on genome organization and stability. In: Charlebois, R.L. (Ed.), Organization of the Prokaryotic Genome. ASM Press, Washington, DC, pp. 109–128.

Jeong, H., Mason, S., Barabasi, A., Oltvai, Z., 2001. Lethality and centrality in protein networks. Nature 411, 41–42.

Kibota, T.T., Lynch, M., 1996. Estimate of the genomic mutation rate deleterious to overall fitness in *E. coli*. Nature 381, 694–696.

Kimura, M., 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge.

Knibbe, C., Beslon, G., Lefort, V., Chaudier, F., Fayard, J.-M., 2005. Self-adaptation of genome size in artificial organisms. In: Capcarrere, M.S., Freitas, A.A., Bentley, P.J., Johnson, C.G., Timmis, J. (Eds.), Advances in Artificial Life, Proceedings of the Eighth European Conference, ECAL 2005. Lecture Notes in Artificial Intelligence, vol. 3630. Springer, Berlin, pp. 423–432.

Layzer, D., 1980. Genetic variation and progressive evolution. Am. Nat. 115, 809–826.

Martin, G., Lenormand, T., 2006. A general multivariate extension of Fisher's geometrical model and the distribution of mutation fitness effects across species. Evolution 60, 893–907.

Morange, M., 2000. Gene function. Life Sci. 323 (12), 1147–1153.

Morgan, T.H., 1926. The Theory of the Gene. Yale University Press, New Haven, CT.

Rocha, E.P.C., 2003. An appraisal of the potential for illegitimate recombination in bacterial genomes and its consequences: from duplications to genome reduction. Genome Res. 13, 1123–1132.

Schuster, P., Swetina, J., 1988. Stationary mutant distributions and evolutionary optimization. Bull. Math. Biol. 50, 635–660.

Van Nimwegen, E., Crutchfield, J.P., Huynen, M., 1999. Neutral evolution of mutational robustness. Proc. Natl Acad. Sci. USA 96, 9716–9720.

Wagner, G., Altenberg, L., 1996. Complex adaptations and evolution of evolvability. Evolution 50, 967–976.

Wagner, G.P., Booth, G., Bagheri-Chaichian, H., 1997. A population genetic theory of canalization. Evolution 51, 329–347.

Whitley, D., 1989. The GENITOR algorithm and selection pressure: why rank-based allocation of reproductive trials is best. In: Schaffer, J.D. (Ed.), Proceedings of the Third International Conference on Genetic Algorithms. Morgan Kaufmann Publishers Inc., San Mateo CA, pp. 116–121.

Wilke, C.O., 2001. Adaptive evolution on neutral networks. Bull. Math. Biol. 63, 715–730.

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

Wright, S., 1968. Evolution and the Genetics of Populations, vol. 1. University of Chicago Press, Chicago.

Wuchty, S., 2004. Evolution and topology in the yeast protein interaction network. Genome Res. 14 (7), 1310–1314.