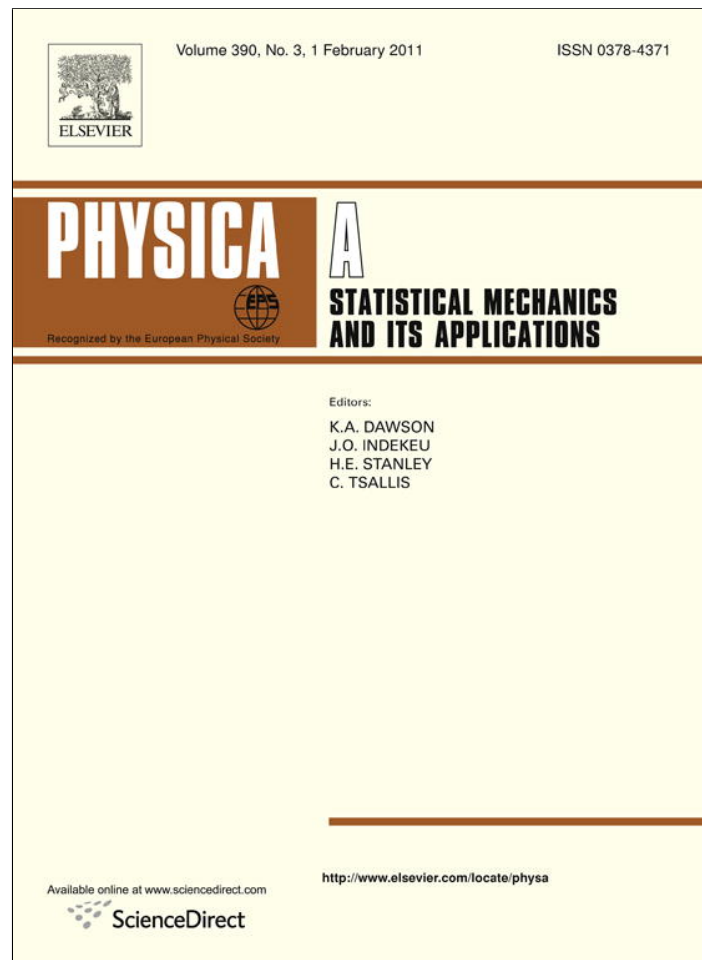


Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.



This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/copyright>



Contents lists available at ScienceDirect

Physica A

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

The zipper effect: Why different positions along the chromosome suffer different selection pressures

P.M.C. de Oliveira*, S. Moss de Oliveira

Instituto de Física, Universidade Federal Fluminense, Av. Litorânea s/n, Boa Viagem, Niterói 24210-340, RJ, Brazil
National Institute of Science and Technology for Complex Systems, Brazil

ARTICLE INFO

Article history:

Received 7 October 2010

Available online 19 October 2010

Keywords:

Evolution

Crossing-over

ABSTRACT

Variability within diploid sexual populations comes from two ingredients: mutations and recombination (or crossing-over). On average, the first introduces genetic defects in offspring genomes, while the second is a mechanism which tends to eliminate them, continuously “cleaning” the population genetic pool from harmful mutations along the generations. Here, we propose that loci near the chromosome tips are more effectively cleaned by the recombination mechanism than loci near the chromosome centre. This result implies that clusters of neighbouring, orchestrated-functioning genes, supposed to be more robust against the effects of genetic mutations, are more likely found near the chromosome centres, while isolated genes are more likely found near the tips. We confirm the tip–centre asymmetry through a simple computer agent-based model. In order to test this effect in reality, we also analyse as an example the particular case of HOX genes distributed along the 24 human chromosomes and verify that indeed, most HOX genes belong to such clustered networks located near the chromosome centres. Accordingly, isolated HOX genes are located closer to the tips.

© 2010 Elsevier B.V. All rights reserved.

1. Introduction

In the past it was believed that each gene corresponds to just one locus, that is, to a single piece of adjacent DNA bases (A, T, G and C). Also, no role was attributed to the intergenic material that separates one gene from the other. Over the last two decades two enormous improvements were obtained concerning the understanding of genetic information storage. First, parts of the intergenic material have been recognised as playing fundamental roles in the regulation of genetic expression. Second, it has also been recognised that some genes are not a single piece of adjacent bases but are formed instead, by a certain number of such pieces, which we will call here sub-pieces, separated from each other. (Biologists normally have special names for each sub-piece, according to its particular function. Here, we are not interested in their specific function, therefore we use the general denomination “sub-pieces”.) Although separated, the sub-pieces forming a gene (or a cluster of interdependent genes) are located at the same region along the chromosome. In general, the expression of these multi-loci genes are strongly related to the regulatory mechanism. Different genetic functions can be obtained by different combinations of the same DNA sub-pieces present at the same chromosome region. The current interpretation is that evolution follows the strategy of giving new uses to old genetic material, by combining them differently, instead of inventing new ones, i.e. some sort of bricolage [1]. Mutations occurring in some of these sub-pieces can also produce slightly different functions for the otherwise same gene. Frequently one finds copies of the same sub-piece in different positions along the chromosome, slightly different from each other due to mutations that occurred in the genome of some founder ancestors. The use of one or the other among these similar sub-pieces produces different genetic functions, a further source of genetic diversity offered to the regulatory system.

* Corresponding author. Tel.: +55 21 2629 5825; fax: +55 21 2629 5887.

E-mail addresses: pmco@if.uff.br (P.M.C. de Oliveira), suzana@if.uff.br (S. Moss de Oliveira).

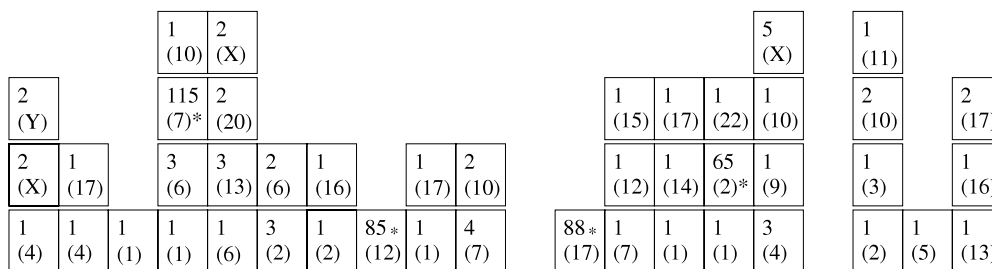


Fig. 1. Along the horizontal direction, each of the 24 human chromosomes were divided into 20 adjacent segments. Each rectangular box displays the number of HOX genes found inside those segments and the number or letter of the corresponding chromosome inside parentheses. Clusters of many orchestrated-functioning HOX genes can be recognised by the large numbers of genes within the same rectangular box: 115 in chromosome (7), 4th segment; 85 in (12), 8th; 88 in (17), 12th; and 65 in (2), 15th. They are concentrated near the chromosome centres. The remainder genes are not clustered and are more likely located near both chromosome tips, only 23 inside the central half, 41 beyond it.

Written for a broad audience, Ref. [2] presents an excellent description of this subject and the recent improvements on its understanding. It also presents a good list of specialised scientific works. A recent interview with Denis Duboule [3], a pioneer of the HOX gene clustering organisation discovery [4], can also be useful. One of the most important concepts recently acquired is the temporal expression of genes controlled by their spatial distribution along the chromosome (see for instance [5–7]), which is just the subject of the present work.

For our purposes here, we will divide genes into two broad classes: single-locus and multi-loci genes. According to our proposal, genes of the first class are preferentially located near the tips of the chromosomes while those of the second class are found preferentially near the centres.

One can understand our reasoning through a simple mechanism, similar to the so-called Hill–Robertson effect [8]. Imagine two heterozygous loci, one Aa and the other bB , separated by some distance along the parent’s genome, where A and B are the wild alleles and a and b are some mutated, less effective alleles which provide some negative selective handicap. A and B belong to one chromosome, whereas a and B belong to the homologous chromosome. If the crossing occurs between these two loci, one gamete presenting only the wild alleles A and B is produced, with both a and b “cleaned” out. On the other hand, if the crossing does not occur between these two loci but at some other position outside the interval separating them, no cleansing happens at all. The larger the distance between these two heterozygous loci Aa and bB along the parent’s genome, the larger is the probability of cleaning both loci on the corresponding gamete. Therefore, a spatial correlation along the genome emerges as a consequence of the crossing mechanism. Two neighbouring heterozygous loci along the parent’s genome, separated by only a few adjacent loci, are very difficult to be cleaned: For that, the crossing position should occur inside the tiny interval separating both loci.

In order to test our hypothesis we analysed the so-called HOX genes for humans. They are now known as responsible for the most primitive and fundamental features of embryogenesis, the regulatory system for lateral symmetry, segmentation, making of members, etc., present in virtually all animals. We took from Ref. [9] the initial and final positions of 417 HOX genes along each of the 24 human chromosomes (1, 2, ..., 22, X and Y). Since chromosomes present different total lengths, we divided each one into 20 adjacent segments, each segment containing 5% of the chromosome’s length. Any HOX-gene reported in Ref. [9] is entirely located inside such a segment. Then we counted how many HOX genes were found in the first segment of each chromosome, then in the second segment, and so on. The resulting histogram is shown in Fig. 1, where one can note that clusters of many orchestrated-functioning HOX genes are more concentrated near the chromosome centres, while isolated ones are near the tips.

Such a result supports our proposed theoretical scenario, that is: (1) Clusters of multi-loci, multi-function genes, for which mutations are more likely to help in the adaptation to the current environment (see, for instance, [10]) tend to be located near the chromosome centres, while single-locus genes, for which mutations rarely improve their genetic functionality, tend to be located near the chromosome tips; (2) Since mutations occur randomly along the whole chromosome, in principle without any preference towards neither the centre nor the tips, another fundamental heredity ingredient should play the evolutionary role of locating some genes preferentially near the chromosome centre or near the tips, according to their genetic functions. The strongest possible candidate is recombination (or crossing-over), a heredity ingredient which can distinguish positions along the chromosome.

The purpose of this work is to provide a convincing argument that this mechanism indeed occurs, by showing how it works. For that, we implemented a simple population dynamics model which considers only Mendelian heredity and Darwinian selection, operating through random point mutations and recombination.

The next section presents the model, followed by the Results and the Conclusions sections.

2. Model

We use here a model already introduced before [11,12], with some minor modifications. Each individual genome consists of two parallel bit-strings of length L , each bit position representing one of the L diploid loci disposed along the individual

“chromosome”. At each locus bits 0 correspond to the wild allele, while bits 1 represent deleterious mutations inherited from the ancestry.

The construction of the haploid gamete is done as follows. First, one parent’s diploid “chromosome” is copied with some random mutations. On average, m positions along each bit-string are randomly tossed, and the corresponding bits are flipped ($0 \leftrightarrow 1$). The mutation rate m is a parameter not necessarily an integer, since it is an average over the whole population. Then, the crossing position is randomly chosen along the diploid “chromosome”. The parallel bit-strings are cut at that position, and the left (right) part of the first (second) bit-string is attached to the right (left) part of the second (first) bit-string, generating two gametes. One of them is chosen at random to be inherited by the offspring. The other parent’s diploid “chromosome” provides the second gamete, constructed by the same rule, to compose the offspring genome. Both parents are also chosen at random for each newborn.

Selection comes from the survival probability of each individual, which depends on its genome. The wild allele, bit 0, is considered dominant. The deleterious allele, bit 1, affects the individual survival only if it is present at the same locus in both bit-strings. The survival probability is given by x^{N+1} , where N counts the number of 11 loci along the individual diploid genome; x is a parameter slightly smaller than unity, which represents the selection pressure. The larger the x is, the weaker is the selection pressure. At each time step, each individual is killed with probability $1 - x^{N+1}$. Individuals with “dirty” genomes (many alleles 1 and consequently many 11 loci) are more likely to die. After this death mechanism, but still at the same time step, the survivors breed in order to restore the population size of 1000 individuals. Extinction may occur if there is less than two survivors at some time step.

3. Results

At the beginning, all 1000 founder individuals carry only 0 alleles. As time evolves and newborns appear, random mutations are introduced and some bits 1 appear in the population genetic pool. Without the selection mechanism, the fraction of 1 or 0 alleles would eventually stabilise around 50%, a mutational meltdown. However, the selection mechanism based on the x^{N+1} survival probability can avoid this meltdown, if the mutation rate is not too high, allowing the fraction of 0 alleles to remain far above 50%.

Indeed, for the case where bit 0 is recessive, a threshold $m_c \approx 1$ was precisely found [12], above which the fraction of bits 1 explosively grows towards 50% and extinction occurs.¹

For the current more interesting case where bit 1 is recessive, however, we have observed that the population can survive until $m \approx 1.2$: for this value, extinction does not occur in some computer realisations and the fraction of bits 1 remains far below 50%, even after an exceedingly high number of generation.

In order to investigate this extinction occurrence, we fixed $m = 1.2$ and followed the evolution of several populations, differing only by the initial random seeds. Fig. 2 shows the evolution of a typical population with $L = 16, 384$ and $x = 0.9$. Extinction occurs for that particular population only after 23,128,835 time steps, when a single individual survives. The large plot shows only the first 2 million time steps, and the inset shows the whole evolution.

The first 300 000 time steps (symbol T_g) corresponds to the transient during which both the 11-homozygosity (fraction of 11 loci—bottom plot) and the heterozygosity (fraction of 01 and 10 loci), averaged over the whole population, almost vanish. During this transient, the survival rate (fraction of deaths per time step) stabilises around 70%. Suddenly, the genetic degeneration starts, both the 11-homozygosity and the heterozygosity grow up. The survival rate continuously decreases until extinction. What really occurs at the precise moment T_g when genetic degeneration starts?

With the same parameters of Fig. 2, Fig. 3 corresponds to another population (which becomes extinct after 18,610,316 time steps), and shows snapshots of the 1-bit distribution along the linear genome, at 2.86, 2.87, 2.88, 2.89, 2.90, 2.91, 3, 3.2 and 5 successive millions of time steps (typewrite reading sense). The horizontal axis (common to all snapshots) displays adjacent loci along the genome. The red (green) points display the fraction of individuals presenting homozygous 11 (heterozygous 01 or 10) genetic state(s) at each particular loci. Before the genetic degeneration starts (top left) both fractions are small along the whole genome. When the genetic degeneration starts (T_g , top middle), the fraction of 1-bit alleles becomes larger within a small region of the genome. This region grows in width (top right and subsequent plots) corresponding to wider and wider plateaux. Genetic degeneration propagates as a wave along the genome, triggered at some particular position, the focus, where it was nucleated by chance. Other realisations show the same behaviour, although the degeneration focus appears at different times as well as different positions along the chromosome.

This somewhat surprising zipper effect can be understood as follows. As a consequence of selection, within a surviving, stable population the 0-bit allele occupies the great majority of the genomes, only a few loci present 1-bits along each individual genome. Therefore, at births, “bad” mutations $0 \rightarrow 1$ are much more likely to occur than “good” mutations $1 \rightarrow 0$, because of the random choice of the specific locus where each mutation is set. Mutations, then, are the main source of genetic degeneration. Generation after generation, genetic stability must be sustained by some repairing mechanisms, the main one being recombination or crossing-over performed on parent’s genomes, which can produce better gametes

¹ It is interesting to note that $m \approx 1$ coincides with the reported number of mutations per human chromosome, if one considers mutations occurring in $\approx 5\%$ of the human DNA which encodes proteins [13–16] (in our simple model, intergenic material is absent). Other animals with more-than-one diploid chromosome per cell also present ≈ 1 mutation per chromosome per offspring.

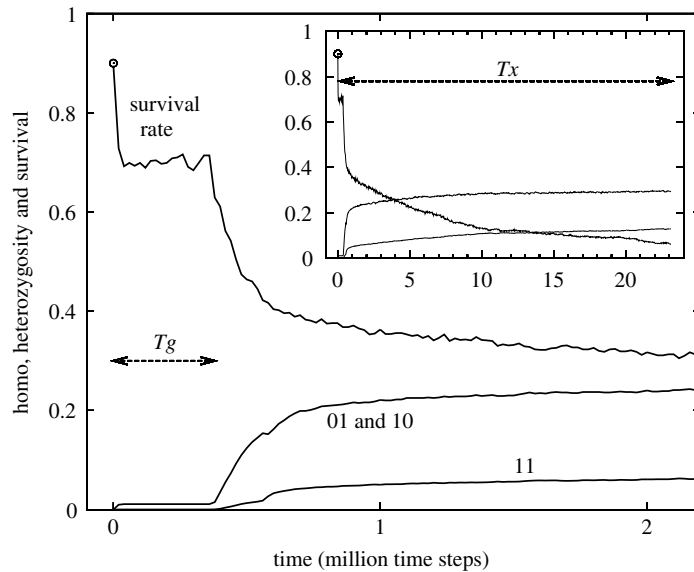


Fig. 2. Evolution of a typical population. After a transient time T_g during which the population still succeeds in keeping a good genetic quality (survival rate $\approx 70\%$), the genetic degeneration suddenly starts and grows continuously until extinction at time T_x . See the main text.

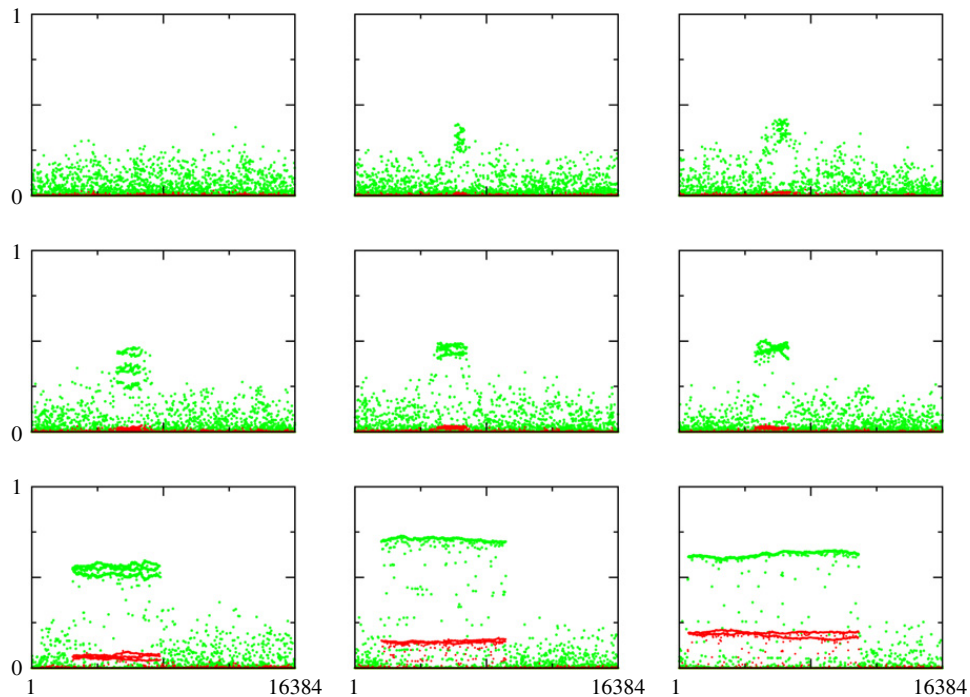


Fig. 3. Another typical population. Genetic degeneration propagates as a wave along the genome (horizontal axis), suddenly triggered at some focus near the chromosome centre (top middle), after $T_g \approx 3$ million time steps in this particular realisation. The vertical axis measures the fraction of heterozygous 01 or 10 individuals at each locus (upper plateau, green symbols) as well as the fraction of homozygous 11 (bottom plateau, red symbols). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(with even less 1-bit alleles than the parent's) to be passed on to lucky offspring. In short, crossing and recombination are the main source for “cleaning” the population genetic pool, counteracting the degeneration produced by mutations.

Two heterozygous loci, one 01 and the other 10, separated by some small distance along the parent's genome are very difficult to be cleaned: For that, the crossing position tossed at random should occur inside the tiny interval separating both loci. That is why the sudden clustered burst of 1-bits which appears around some stochastic position and time becomes fixed on the population genetic pool from then on. Moreover, its left and right wave-fronts move away from each other along the chromosomes, leading eventually to extinction as shown in Fig. 3. Anyway, loci near the extremities are more protected against these fronts, which thus propagate towards the extremities with decreasing speed.

Different realisations with the same genome length L present different transient times T_g when the genetic degeneration suddenly starts (Fig. 2 and top middle frame of Fig. 3). Although difficult to measure because many long computer runs are

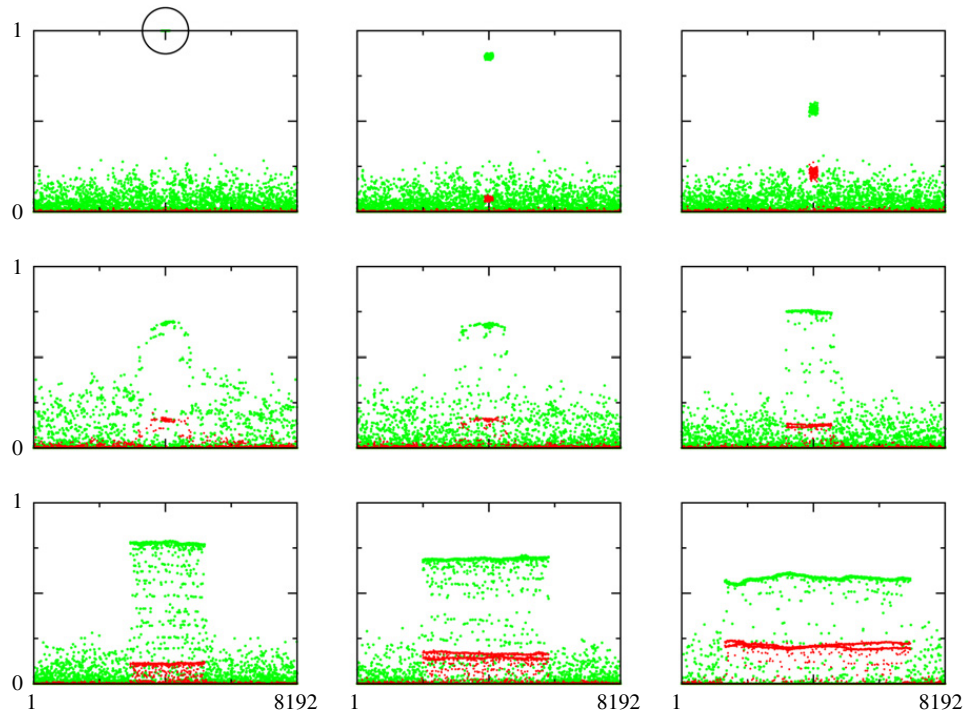


Fig. 4. The same as Fig. 3, for a shorter genome length $L = 8192$. Now, the ignition of the genetic degeneration was artificially induced after one million time steps (top left) by setting heterozygous some few loci near the genome centre (circle).

necessary, the typical T_g seems to depend on the genome length L according to a curious behaviour. For $L = 16,384$, T_g fluctuates more often in between 1 million and 10 million time steps (the particular realisation of Fig. 2 corresponds to an unlucky population which gets the genetic degeneration too early). For the larger genome length $L = 32,768$, the few realisations we tested present T_g shorter than one million time steps, for the same set of parameters, indicating that T_g decreases for increasing L . Indeed, smaller genome lengths $L = 8192, 4096$ and 2048 did not spontaneously show the expected genetic degeneration after 300 million time steps, when we stop our computer runs. Near the left and right extremities of the linear genome the degeneration-triggering mechanism is inhibited, the degeneration focus is likely to appear far from the extremities. Indeed, the genetic degeneration can only be observed within a reasonable time scale for large enough genome lengths, where the genome's central region is far enough from the tips to become free of the inhibition effect (i.e. the genome length is larger than the inhibition penetration length). The majority of our runs for $L = 16,384$ show mutational meltdown starting near the genome centre as in Fig. 3. No realisation where the distance focus-extremity is less than 2 thousand loci was found.

The same crossing/recombination mechanism which inhibits the triggering of the degenerative process near the extremities is also responsible for a similar effect: Loci positioned near the extremities present less defective genes, on average. They are “cleaned” faster and therefore more robust against degeneration. This effect is also responsible for the survival rate behaviour shown in Fig. 2: It begins to decay very fast just after degeneration starts, when only a small region of the genome far from the extremities is already degenerated. The wavefront propagates fast at the beginning, but continuously cools thereafter.

Instead of waiting for a long time until genetic degeneration spontaneously starts, we decide to induce it artificially for $L = 8192$ as in Fig. 4. The parameters are the same as before: $m = 1.2$ and $x = 0.9$. First, the population evolves normally during the first one million time steps, when the genetic degeneration was not yet spontaneously triggered. Then, we toss 128 among the 256 loci located around the genome centre, and randomly set these loci heterozygous 01 or 10 for all individuals (top left). The subsequent plots show the genetic distribution along the genome after 1, 10, 100, one, ten and a hundred thousand, one and ten million time steps (typewrite reading sense). The same behaviour is observed for other genome lengths as well as other sets of parameters ($m > 1$).

For m slightly above 1, the damaged segment stops to grow in some realisations, the degenerated plateau (similar to Fig. 3 or 4) become stuck near the centre and do not touch the genome extremities. In other realisations the plateaux even retract till they vanish, after some initial growth. This effect also seems to be related to the already quoted inhibition near the genome extremities. Roughly, degeneration is created near the genome centre, and is blocked by genetic restoration at the extremities. Which effect wins this dispute, and whether the population survives or not depend on the average number m of mutations at birth. Also, large diploid genomes, for which the extremities play a little role, are relatively less protected against genetic degeneration. The inhibition effect can also be controlled by more or less crossings and recombinations realised for each gamete (all results shown here correspond to just one crossing/recombination per gamete, but we have tested a few other cases).

We are driven to the question: Why do humans present 23 chromosome pairs instead of a single, long one? Another similar question: Why diploid chromosomes are linear, and not circular like bacteria? Yet another: Why genetic material seems to be more robust against degeneration near the chromosome tips, compared to chromosome centres? Perhaps the results obtained from our simple model can help the understanding of these questions.

Note that no spatial correlation at all is introduced along the genome by the model definition, all genes are considered equivalent. The observed correlation naturally emerges, degeneration near the genome centre and restoration near the extremities, due to crossings and recombinations realised in gamete production.

Other curious spatial correlation effects induced by crossing/recombination are studied by geneticist Stanislaw Cebart and his group [17]. They observed the so-called complementarity: almost all alive individuals present the same heterozygous segment (for example . . . 0010110010111 . . . and . . . 1101001101000 . . . , within the bit-string model) at the chromosome central region. This gives no handicap at all for the individual, due to recessiveness. Moreover, a couple of such individuals can breed lucky offspring, provided no crossing/recombination event occurs inside the heterozygous segment. Sharing the same central heterozygous pattern with the parents (and almost the whole population), these lucky offspring will again hide their 1 alleles behind recessiveness, avoiding handicaps in spite of the large number of 1-bits. Distributed on a huge geographic region, with limited search for sexual partners, this behaviour eventually leads to speciation [17], different species can be recognised by their central chromosome region. Individuals belonging to different species cannot breed viable offspring. These works [17] also consider different crossing ratios, the larger it is the more effective is the cleansing mechanism.

An entropic interpretation is also interesting. At the beginning, the genetic variability among the alive members of the population is low, all genomes have mostly 0 alleles. As generations pass, 1 alleles are introduced by random mutations and the variability (or entropy) increases. Later, when the whole population reaches the same heterozygous pattern at the chromosome central region, variability decreases again. The heterozygous central segment becomes stuck, an absorbing state within this segment no longer useful for evolution.

4. Conclusions

Based on a simple argument we predict that the crossing is more effective in repairing diploid DNA degeneration near the chromosome tips than the central region, as generations pass. The argument follows in short. One locus along the diploid chromosome presents one damaged allele on the first DNA double chain but the corresponding allele on the second (homologous) double chain is intact. Another locus presents the inverse situation, one intact allele on the first double chain, a damaged one on the second. During meiosis, if the crossing position occurs between these two loci, one gamete will be produced where both damages are simultaneously “cleaned”, no damage on both loci. On the other hand, the probability of having crossing between these two loci is larger the larger is the distance from each other. Therefore, the repairing mechanism works better for loci located the largest possible distance from each other, i.e. near the chromosome tips.

Computer simulations of a simple bit-string model confirms this effect, and shows how a sudden degeneration nucleated on some small region of the diploid chromosome can propagate as a wave towards the tips, as the opening of a zipper started near the centre.

An inspection of the HOX gene locations along human chromosomes is compatible with our findings. HOX genes regulate fundamental embryogenetic functions, from the most simple to the most complex animal. Therefore, they are supposed to be robust against damages. Indeed, 64 simple HOX genes formed each one by a single DNA segment of adjacent basis are likely to be found near the human chromosome tips than in the central region. On the other hand, 3 complexes of HOX genes formed by many different DNA segments slightly separated from each other are found near the central part of human chromosomes 2, 12 and 17. Each of these complexes work together, as an orchestra with many possible redundancies. Mutations there cannot be considered as damages to be repaired, on the contrary they can be very useful in offering a high degree of diversity through slightly distinct genetic functions of slightly modified DNA pieces forming a single gene. Only a fourth of such complexes is located closer to one tip than to the central region of human chromosome 7, but anyway not so near the tip (note that chromosomes 2 and 7 are larger than 12 and 17, and see the positions of their HOX gene clusters in Fig. 1, symbols *).

References

- [1] D. Duboule, A.S. Wilkins, *Trends Genet.* 14 (1998) 54.
- [2] S.B. Carroll, *Endless Forms Most Beautiful (The New Science of Evo Devo and the Making of the Animal Kingdom)*, W.W. Norton, NY, 2005.
- [3] D. Duboule, *Int. J. Dev. Biol.* 53 (2009) 717.
- [4] D. Duboule, P. Doll, *EMBO J.* 8 (1989) 1497;
A. Graham, N. Papalapov, R. Krumlauf, *Cell* 57 (1989) 367.
- [5] R. Krumlauf, *Cell* 78 (1994) 191.
- [6] M. Wijgerde, et al., *Nature* 377 (1995) 209.
- [7] N. Soshnikova, D. Duboule, *Epigenetics* 4 (2009) 537.
- [8] W.G. Hill, A. Robertson, *Genet. Res.* 8 (1966) 269.
- [9] <http://genome.ucsc.edu/cgi-bin/hgTracks?hgtdummyEnterButton.x=0&hgtdummyEnterButton.y=0&hgsid=132706168&clade=mammal&org=Human&db=hg18&position=hox&pix=800>.

- [10] D.J. Earl, M.W. Deem, *Proc. Natl. Acad. Sci. USA* 101 (2004) 11531;
H.S. Rothenfluh, A.J. Gibbs, R.V. Blanden, E.J. Steele, *Proc. Natl. Acad. Sci. USA* 91 (1994) 12163;
M. Yamada, S. Hudson, O. Tournay, S. Bittenbender, S.S. Shane, B. Lange, Y. Tsujimoto, A.J. Caton, G. Rovera, *Proc. Natl. Acad. Sci. USA* 86 (1989) 5123.
- [11] P.M.C. de Oliveira, *Theory Biosci.* 120 (2001) 1. www.arXiv.org/abs/cond-mat/0101170;
P.M.C. de Oliveira, *Physica A* 306 (2002) 351. www.arXiv.org/abs/cond-mat/0108234;
D. Stauffer, S. Moss de Oliveira, P.M.C. de Oliveira, J.S. Sá Martins, *Biology, Sociology, Geology by Computational Physicists*, Elsevier, Amsterdam, ISBN: 0-444-52146-1, 2006.
- [12] P.M.C. de Oliveira, S. Moss de Oliveira, D. Stauffer, S. Cebrat, A. Pekalski, *Eur. Phys. J. B* 63 (2008) 245. www.arXiv.org/abs/q-bio/07101357.
- [13] F.M. Salzano, *An. Acad. Brasil. Ciênc.* 77 (2005) 627.
- [14] A.J. Fry, J.J. Wernegreen, *Gene* 355 (2005) 1.
- [15] N.A. Moran, A. Mira, *Genome Biol.* 2 (2001) 0054.1.
- [16] J.W. Drake, *Proc. Natl. Acad. Sci. USA* 88 (1991) 7160.
- [17] D. Mackiewicz, S. Cebrat, 2009. www.arXiv.org/abs/0901.1465;
M. Zawiarta, W. Waga, D. Mackiewicz, P. Biecek, S. Cebrat, *Internat. J. Modern Phys. C* 19 (2008) 917;
M. Zawiarta, P. Biecek, W. Waga, S. Cebrat, *Theory Biosci.* 125 (2007) 123.