



ELSEVIER

Physica A 273 (1999) 70–74

PHYSICA A

www.elsevier.com/locate/physa

Studying DNA evolution through successive file editions

Paulo Murilo Castro de Oliveira

*Instituto de Física, Universidade Federal Fluminense, av. Litorânea s/n, 24210-340,
Boa Viagem, Niterói RJ, Brazil*

Received 11 June 1999

Abstract

We propose an experimental way to test theories concerning DNA evolution mechanisms, through computer simulations. © 1999 Elsevier Science B.V. All rights reserved.

PACS: 07.05.Tp; 05.10.-a

Keywords: Long-range correlation; DNA evolution; Computer simulation

Long-range correlations (LRC) along DNA sequences corresponding to different species were measured by different groups [1–10]. The successive basis A, T, G and C are translated into some numerical steps, for instance, A and T corresponding to +1, G and C to –1, and a statistical treatment is then performed on the numerical series obtained by adding these successive steps. The results are controversial, some authors [1–7] claim that no LRC appear, while others [8–10] present data showing them. The overall observations of Refs. [8–10] were: (1) correlations indeed exist when the complete sequence is read, including introns; (2) by reading only exons, i.e. skipping introns during the reading procedure, correlations do not appear; (3) the “more evolved” the species under study is, the higher is the degree of such correlations, and also the higher is the degree of introns’ content. Exons are the parts of DNA sequences known to code for some protein synthesis, while introns are portions located in between exons with no apparent biological function (I am not interested here in the polemic question about which is the biological function of introns). The first obvious conclusion is that biological evolution would gradually introduce both correlations and introns along DNA, after successive generations.

E-mail address: pmco@if.uff.br (P.M.C. de Oliveira)

The genetic material of primitive organisms as algae is coded by small DNA sequences formed almost exclusively by exons. Complex organisms as humans present enormous DNA sequences formed mostly by introns (only 5–10% of the human genetic code corresponds to exons). From these observations a beautiful theory [8–10] emerges about DNA evolution mechanism. As complex organisms evolve from primitive ones, the quantity of genetic material gradually increases during this slow process. During reproduction, DNA sequences are replicated with errors (mutations) of which some of them remain within the populations after many generations. These errors can be point mutations (a single basis is replaced by another one in the offspring genetic material), deletions (an entire piece of DNA is missing in the offspring genetic material), and insertions (some DNA piece is copied twice, in two different locations of the offspring genetic material). Thus, it would be possible to find the same exon twice, in different locations, along the same DNA sequence, coding for the same protein. However, further mutations (occurred some generations after the initial insertion which leads to the double exon) are likely to damage one of these copies, transforming it into a “useless” part. There is no selection pressure against this damage, because the other copy remains intact. According to the quoted theory [8–10], these damaged DNA parts are the currently observed introns. Indeed, they are: (1) more frequently observed in more evolved organisms; (2) not identical, but similar to other (not damaged) DNA parts, leading to the observed long-range correlations.

Unfortunately, this beautiful theory cannot be tested by direct biological experiments. One cannot easily re-make the historical evolution path of a species. However, the dynamical mechanism claimed as responsible for the appearance of introns and the consequent long-range correlations could be tested in other, non-biological experiments. A possible test for such mechanisms is the successive file editions within a computer diskette [11–13]. In these references, a particular operating system, namely DOS, is used in order to perform successive editions into a set of random files stored in a diskette. Each new edition consists in changing some random bits (point mutations), deleting some material (deletions), or including some new randomly chosen material (insertions). The operating system, however, does not store the new file version exactly in the same place where the old version is already written: some pieces of the old version could remain stored on the diskette surface, as introns, being simply marked in a table as empty for future use. After each edition, the long-range correlations along the whole diskette are measured, and the result is that: (1) these correlations indeed exist when the whole diskette surface is read, including parts not currently belonging to files (introns); (2) there is no correlation at all when only the actual files are read, skipping parts not currently in use (only exons); (3) as more and more successive editions are performed, the higher is the degree of correlations.

LRC can be detected by measuring the fluctuation $F(l)$ as a function of the distance l along the DNA sequence [8–10], or along the diskette sequence of bits, or along any series of data. By plotting this function on a double logarithm scale, the result is a straight line (at least for large values of l). Would the slope α of this line be equal to $\frac{1}{2}$, as a random walk, one has no LRC at all. Indeed, LRC correspond to $\alpha \neq \frac{1}{2}$, and

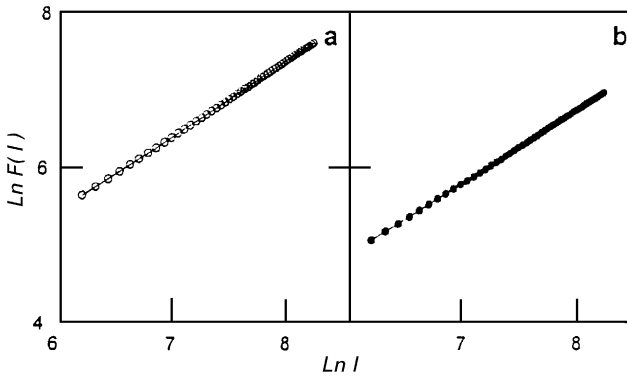


Fig. 1. Fluctuations measured in a real diskette, by reading: (a) only parts belonging to current files (skipping introns); or (b) the whole diskette (including introns).

the deviation between the value of α and $\frac{1}{2}$ measures the degree of LRC. For the DOS diskette system [13], taking an ordinary diskette (containing some C-source codes) already used many times to edit and save the files, the fluctuation function is shown in Fig 1. Part a considers only the currently stored files, skipping reading the inter-file information (introns). Part b considers the whole diskette. In this case, both values for α are near to unity, once the redundance of C programming language is very high. Alternatively, by starting from files previously prepared with random sequence of bits, and implementing random editions including insertions and deletions, one can observe the dynamical evolution of α : it gradually increases (starting from $\alpha = \frac{1}{2}$) as more and more editions are implemented, for the case where all the diskette information is read (including introns). By reading only the parts currently in use by some file (skipping introns in the reading procedure), however, the value of α remains always near $\frac{1}{2}$ [11,12].

In the present work, instead of a real computer operating system with its particular rules for storing informations, we propose to perform the tests according to real (theoretical) biological rules. Some biologist (not myself, of course, a physicist ignorant of biology) could figure out which were the actual rules adopted by mother Nature in what concerns DNA evolution. Given these rules, they can be programmed in order to perform the evolution of a primitive DNA sequence on the computer, and the result after many iterative applications of the same rules can be compared with the observations on real DNA sequences. The DNA is represented by a long sequence of bits 0 or 1, which is gradually modified according to the rules under test. In each step, another sequence of bits is created, storing the partial results arisen from the application of the various rules. At the end, the new sequence is copied into the original one, before starting to apply again the rules (for simple rules, one sequence edited onto itself could be enough). The sequences can store the “genetic” information of all individuals of the current population, and thus some extra rules concerning their reproductive behaviour must be programmed as well. In this case, many redundant information (the same gene

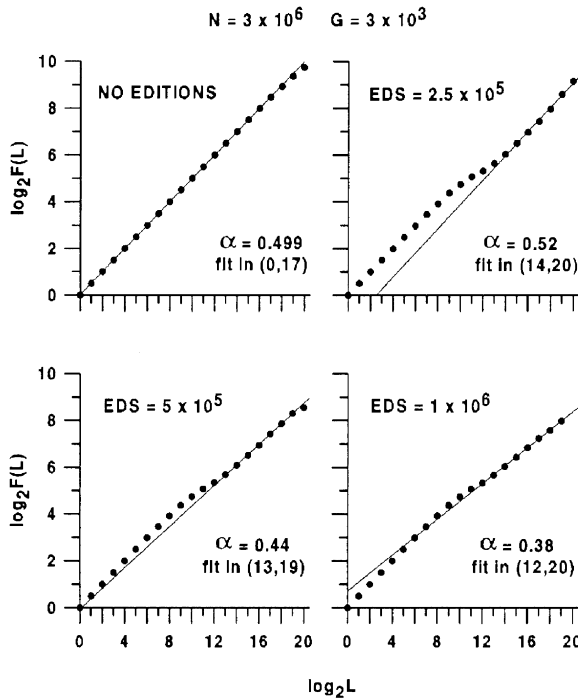


Fig. 2. Illustrative example of successive editions of a initially random sequence of N bits. A fixed “gene” with G bits is copied again and again into random positions along the sequence. As time goes by, the slope α deviates from the initial value $\frac{1}{2}$, for length scales larger than G , indicating the appearance of long-range correlations.

for different individuals) are kept during the computer processing. In order to perform a fast test, the sequences can be thought as storing the current genetic pool of the whole populations, at each step. The degree of long range correlations can be measured after each step, and its time evolution analysed, whereas only the final (or current) degree of correlations along real DNA sequences can be measured.

Three simple C-programs (available at www.if.uff.br/~pmco/dna) can guide the reader in doing such tests. As an example, they use a very simple, artificial rule: a previously fixed sub-sequence (a fixed “gene”) is copied onto a random position along the main sequence, overwriting whatever was there. At the next step, the same sub-sequence is copied again onto another random position, and so on. The first program DNA.C uses only the first bit of each 32-bit word $DNA[i]$, where i stands for the bit position along the DNA chain. The second program DNA32.C performs the same thing for 32 different chains, in parallel: running it is the same as running 32 times DNA.C, taking the average fluctuation function $F(L)$ at the end. The last program DNABIT.C uses only one chain, as DNA.C, but stores 32 bits along each 32-bit word, saving memory by a factor of 32 [14]. By using DNA32.C, we got the plots shown in Fig. 2, for increasing number of edition steps. One can note that LRC are indeed introduced during the time

evolution. In this case, the length of the whole bit sequence is $N = 3 \times 10^6$, and the fixed “gene” length is $G = 3 \times 10^3$. Accordingly, LRC appear only for scales larger than $\log_2(G) \approx 8$. These results are only illustrative, and do not deserve further analysis. On the contrary, the purpose is to replace this fixed “gene”, naive rule by other more realistic ones the reader would program by him(her)self on routine evolve().

Acknowledgements

I am grateful to Gilney Zebende who produced Fig. 1, and Thadeu Penna for a critical reading of the manuscript.

References

- [1] M.Ya. Azbel', Y. Kantor, L. Verkhil, A. Vilenkin, *Biopolymers* 21 (1982) 1687.
- [2] M.Ya. Azbel', *Phys. Rev. Lett.* 75 (1995) 168.
- [3] W. Li, K. Kaneko, *Europhys. Lett.* 17 (1992) 655.
- [4] R. Voss, *Phys. Rev. Lett.* 68 (1992) 3805.
- [5] S. Cebrat, M.R. Dudek, *Eur. Phys. J. B* 3 (1998) 271.
- [6] S. Cebrat, M.R. Dudek, P. Mackiewicz, *Theor. Biosci.* 117 (1998) 78.
- [7] M.S. Vieira, COND-MAT 9905074.
- [8] C.-K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons, H.E. Stanley, *Nature* 356 (1992) 168.
- [9] S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.-K. Peng, H.E. Stanley, M.H.R. Stanley, M. Simons, *Biophys. J.* 65 (1993) 2675.
- [10] H.E. Stanley, S.V. Buldyrev, A.L. Goldberger, Z.D. Goldberger, S. Havlin, R.N. Mantegna, S.M. Ossadnik, C.-K. Peng, M. Simons, *Physica A* 205 (1994) 214.
- [11] G.F. Zebende, T.J.P. Penna, P.M.C. de Oliveira, *Phys. Rev. E* 57 (1998) 3311.
- [12] G.F. Zebende, T.J.P. Penna, P.M.C. de Oliveira, *Physica A* 257 (1998) 136.
- [13] G.F. Zebende, Ph.D. Thesis, Universidade Federal Fluminense, 1999.
- [14] For similar multi-bit tricks, see P.M.C. de Oliveira, *Computing Boolean Statistical Models*, World Scientific, Singapore, 1991 ISBN 981-02-0238-5.