ELSEVIER

Physica A 249 (1998) 594–599

PHYSICA A

Model of unequal chromosomal crossing over in DNA sequences [1]

Nikolay V. Dokholyan [a,*], Sergey V. Buldyrev [a], Shlomo Havlin [b], H. Eugene Stanley [a]

[a] *Center for Polymer Studies, Physics Department, Boston University, Boston, MA 02215, USA*
[b] *Gonda-Goldschmied Center and Department of Physics, Bar-Ilan University, Ramat Gan, 52900, Israel*

Abstract

It is known that some dimeric tandem repeats (DTR) are very abundant in *noncoding* DNA. We find that certain DTR length distribution functions in *noncoding* DNA can be fit by a power law function. We analyze a simplified model of unequal chromosomal crossing over and find that it produces a stable power law length distribution function, with the exponent $\mu = 2$. Although the exponent predicted by this model differs from those observed in nature, we argue that the biophysical process underlying this model provides the major contribution to the DTR length distribution function.

PACS: 87.10.+e; 05.20.−y; 82.20.Hf
Keywords: DNA; Mutations; Repeats

1. Introduction

The genetic information of organisms is stored in DNA which is a sequence of four different bases: adenine A, guanine G (purines), cytosin C, and thymine T (pyrimidines) [1]. Each DNA molecule is packaged in a chromosome, which varies in length from 10^5 base pairs (bp) in yeast to 10^8 bp in human. Roughly 5% of human DNA is *coding*, i.e. is translated into protein by various combinations of three nucleotides. Although the rest of DNA is *noncoding*, some regions are known to be involved in various regulatory processes. The statistical properties of *coding* DNA are different from *noncoding* DNA. For example, simple sequence repeats (SSR) are very abundant in *noncoding* DNA, but are rare in *coding* DNA. SSR can be represented as $(X_1X_2 \ldots X_N)_\ell$, where $X_{i=1,\ldots,N}$ is one of A, C, G or T, and ℓ is the number of copies, which can, in

* Corresponding author.
[1] This work is supported by NIH-HGP.

PII S0378-4371(97)00520-7

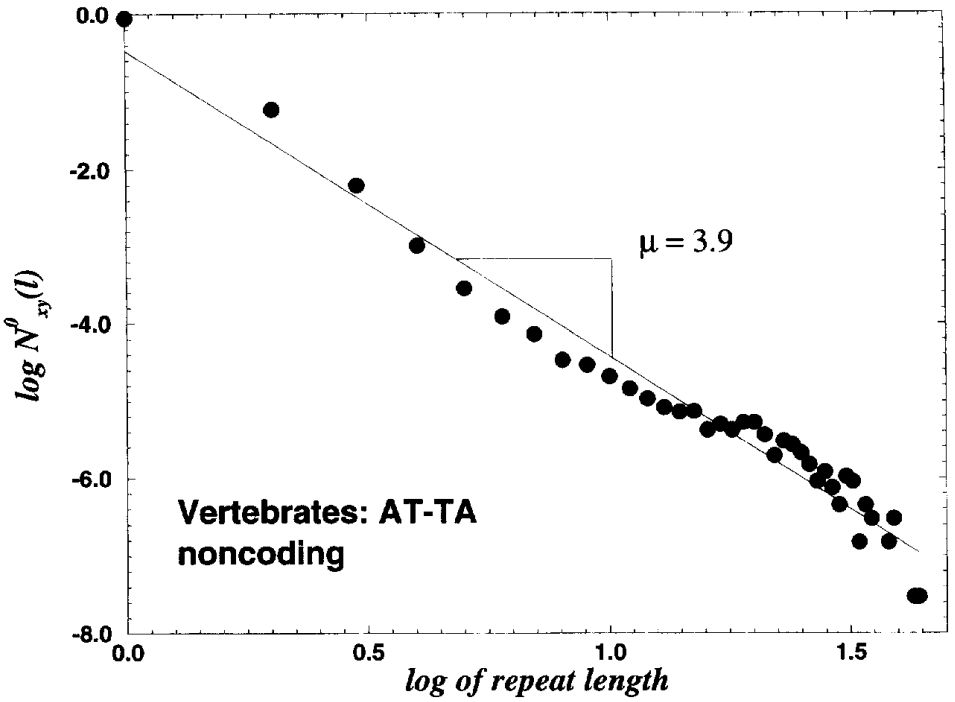

Fig. 1. Double-logarithmic plot of $N^0_{xy}(l)$ for all *noncoding* vertebrate sequences in GenBank (release 96.0); as an example we show the combined results for $AA-TT$. Note that the data can be fit by a power law function, so Eq. (4) holds. We find that the exponent $\mu = 3.9$ for fitting range $l \in [1,44]$, with confidence value $R = 0.98$. The difference of length distribution functions for *noncoding* DNA and uncorrelated random sequence is dramatic; one can observe a DTR of length of 48 dimers in *noncoding* DNA (with probability, roughly, $p \approx 10^{-7}$), while it is 49 orders of magnitude less probable to find such DTR in uncorrelated random sequence (where all bases have equal concentrations) ($p \approx 10^{-56}$)!

some cases, be of the order of 100. SSR are of considerable practical and theoretical interest due to their high polymorphism [2], i.e. the length of SSR at the same DNA locus can vary from individual to individual. This property of SSR is used in genetic fingerprinting.

The expansion of SSR in DNA sequences plays an important role in genome organization and evolution and is associated with many genetic diseases (e.g. see [3–5]). It was recently discovered [6,7] that the length distribution of a particular type of simple repeats, dimeric tandem repeats (DTR), can be fit by a power law function in *noncoding* DNA (see Fig. 1). For vertebrate DNA these DTR include the following: $(AA)_l$, $(CC)_l$, $(GA)_l$, $(AG)_l$, $(GG)_l$, $(TA)_l$, $(AT)_l$, $(TC)_l$, $(CT)_l$ and $(TT)_l$. The remaining of DTR ($(AC)_l$, $(CA)_l$, $(CG)_l$, $(GC)_l$, $(TG)_l$ and $(GT)_l$) behave differently (for details see [6]). To stress the importance of the observed power law DTR length distribution functions, one can estimate that in the case of uncorrelated or short-range correlated random sequence, the length distributions of these DTR would be exponential and long repeats (with $l > 7$) would not be observed in present database (see caption to Fig. 1). The abundance of DTR in noncoding DNA raises a question about the nontrivial biophysical processes which lead to such distributions in the course of

the genome evolution. The difference in statistical properties of *coding* and *noncoding* regions might be a consequence of the fact that less errors are allowed in *coding* DNA than in *noncoding DNA*. Even a single mutation in the *coding* region can lead to the extinction of organisms. For example, the insertion or deletion of a group of nucleotides in *coding* DNA may result in a shift of the reading frame causing misinterpretation of genetic information.

In this paper we conjecture that the DTR length distribution functions follow a power law function due to a specific type of mutation resulting from the unequal crossing over of chromosomes which can occur during cell division (discussed in the next section). We describe a model, based on unequal crossing over [1], which produces power law distributions with exponent $\mu = 2$. We also discuss the role of other biological processes in the evolution of repeat length distribution.

2. Model of unequal chromosomal crossing over

The crossing over of two homologous chromosomes (paternal and maternal) during meiosis is one of the major contributions to the reassortment of genetic information. During meiosis the paired chromosomes can both break due to the tension resulting from the mutual attraction of these chromosomes. Further, the broken ends can be rejoined to the original parental chromosomes or can cross over to join the homologous parental chromosomes. It is sometimes possible that breakage occurs unequally in the two chromosomes. In the case that cross over occurs with such unequal breakage, each parental chromosome changes in length, one becomes longer, while the other becomes shorter. We base our model on this mechanism of unequal chromosomal crossing over, which is defined as follows:

Model. Consider a segment with a DTR of length ℓ (see Fig. 2). We define unequal crossing over to be when a segment of a repeat of length x on chromosome **A** (maternal) joins a segment of a repeat of length $\ell - y$ on chromosome **B** (paternal). As a result, we obtain two repeats of length $\ell - x + y$ (on chromosome **A**) and $\ell + x - y$ (on chromosome **B**), i.e. the chromosome lengths change as follows:

$$\ell_{\mathbf{A}} \to \ell'_{\mathbf{A}} \equiv \ell_{\mathbf{A}} \cdot (1 - \xi_x + \xi_y) \quad \text{or} \quad \ell_{\mathbf{B}} \to \ell'_{\mathbf{B}} \equiv \ell_{\mathbf{B}} \cdot (1 + \xi_x - \xi_y), \tag{1}$$

where $\xi_x \equiv x/\ell$ and $\xi_y \equiv y/\ell$ range from 0 to 1.

Without loss of generality,[2] we follow the evolution of chromosome **A**. Eq. (1) can be rewritten in terms of one parameter $r \equiv 1 - \xi_x + \xi_y$, which ranges from 0 to 2: $\ell \to \ell' = \ell \cdot r$. Thus, if $0 < r < 1$, the repeat shrinks in length, while if $1 < r < 2$, the repeat grows.

[2] Note that in the mathematical sense, chromosomes **A** and **B** are equivalent upon the transformation $x \leftrightarrow y$.

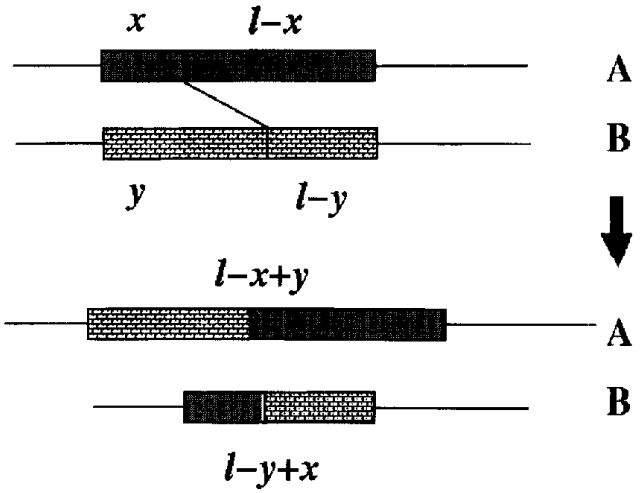

Fig. 2. Crossing over of two homologous chromosomes during the meiosis. Due to unequal crossing over, the part which contains DTR of maternal chromosome **A** of length x joins the part which contains the same type of DTR in the paternal chromosome **B** of length $\ell - y$. As a result, the length of the chromosome **A** changes to $\ell - x + y$. Similarly, the corresponding part of maternal chromosome **A** of length $\ell - x$ joins the same type of DTR in the paternal chromosome **B** of length y. As a result, the length of the chromosome **B** changes to $\ell + x - y$.

To derive the probability distribution function $C(r)$ of variable r we need to integrate over all possible configurations of ξ_x and ξ_y to get a particular value of r:

$$C(r) = \int_0^1 d\xi_x f_1(\xi_x) \int_0^1 d\xi_y f_2(\xi_y) \cdot \delta(r - 1 - \xi_x + \xi_y) . \tag{2}$$

where f_1 and f_2 are distribution functions of ξ_x and ξ_y.

After t steps of evolution the length of the repeat is given by $\ell_t = \ell_0 \cdot \prod_{i=1}^{i=t} r_i$. Such a "random multiplicative process" in many cases leads in the long time limit ($t \to \infty$) to a stable distribution of repeat length $P(\ell)$. To avoid extinction of repeats, one can set the probability of extinction to zero.

If we change variables to $z_t \equiv \ln \ell_t$, the dynamics of z_t becomes a random diffusion process in semi-infinite space $z > 0$ (due to the fact that $\ell \geqslant 1$) with a reflecting wall at $z = 0$ and an attractive potential. A classical example of such a process is Brownian motion in a potential field, which leads in the continuum limit to an exponential decrease of the atmosphere's density. Thus, in this case, the probability distribution function for the final outcome of our process can be written as

$$\bar{P}(z) \sim e^{-kz} , \tag{3}$$

where k is some constant. After transforming back to our original variables, Eq. (3) can be rewritten in the form of power law with $\mu = k + 1$:

$$P(\ell) \sim \ell^{-\mu} . \tag{4}$$

To derive this result, we write the master equation for the z-variable written in the continuous limit:

$$\frac{d\bar{P}(z,t)}{dt} = \int_{-\infty}^{+\infty} \bar{P}(z-x,t) \cdot C(x)\,dx - \bar{P}(z,t)\,. \tag{5}$$

If we assume that $\bar{P}(z,t)$ has a time-independent solution in the form of Eq. (3), then we obtain an additional condition on $C(r)$ in terms of original variable r:

$$\int_0^2 C(r) r^{\mu-1}\,dr = 1\,. \tag{6}$$

It can be shown that for any two distributions f_1 and f_2 with equal first moments Eqs. (2) and (6) yield two solutions $\mu = 1$ and $\mu = 2$. The first solution does not produce a normalizable distribution, so we can disregard it.

3. Discussion

Our model is able to produce power law behavior of DTR length distribution functions with $\mu = 2$. However, for real DNA sequences μ ranges from 2.5 to 4.5 for various taxonomic classes and various DTR. This difference can be attributed to processes, in addition to unequal crossover, that take place in nature.

(i) One possible scenario for $\mu > 2$ can be due to the selective pressure against fixations of the longer repeats in the population, so that the probability of chromosomal growth is larger than that of chromosomal shrinkage. Thus, the function $C(r)$ becomes asymmetric and skewed to the left from the symmetry point $r = 1$. In this case, the solution to Eq. (6) yields power law function with exponent $\mu > 2$. For example, if we take function $C(r)$, asymmetric relative to the point $r = 1$:

$$C(r) = \begin{cases} \frac{3}{4}, & 0 < r \leqslant 1, \\ \frac{1}{4}, & 1 < r \leqslant 2, \\ 0, & r \leqslant 0 \quad \text{or} \quad r > 2, \end{cases} \tag{7}$$

then the solution of the Eq. (6) yields power law function DTR length distribution function $P(\ell)$ with the exponent $\mu \approx 3.7$. This example shows that with slight asymmetry of function $C(r)$ one can achieve any exponent $\mu > 2$ by varying the skewness of the function $C(r)$.

(ii) We suspect, however, that the higher observed exponent μ comes from point substitutions of one nucleotide by another, which are more likely to target longer repeats [8]. Such mutations would destroy long repeats and create more shorter repeats. This would result in a decrease of values of the DTR length distribution functions at the tail of distributions and, therefore, an increase of the exponent μ.

4. Conclusion

We find that a simple mathematical description of unequal chromosomal crossing over is capable of reproducing power law behavior of DTR length distribution functions. We also find that the only restriction one should impose on the probability distribution functions of "target variables" ξ_x and ξ_y is the equality of their mathematical expectations. Taking into account other evolutionary processes, such as point substitutions or slipped strand misspairing, we believe it might be possible to reproduce the exponents observed in nature. This work is currently under investigation.

References

[1] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, J.D. Watson, Molecular Biology of the Cell, Garland Publishing, New York, 1994.
[2] A.M. Bowcock, A. Ruiz-Linares, J. Tomfohrde, E. Minch, J.R. Kidd, L.L. Cavalli-Sforza, Nature 368 (1994) 455.
[3] Huntington's Disease Collaborative Research Group, Cell 72 (1993) 971.
[4] E. Kremer, M. Pritchard, M. Lynch, S. Yu, K. Holman, E. Baker, S.T. Warren, D. Schlessinger, G.R. Sutherland, R.I. Richards, Science 252 (1991) 1711.
[5] Y. Ionov, M.A. Peinado, S. Malkhosyan, D. Shibata, M. Perucho, Nature 363 (1993) 558.
[6] N.V. Dokholyan, S.V. Buldyrev, S. Havlin, H.E. Stanley, J. Mol. Evol. (1997), submitted.
[7] N.V. Dokholyan, S.V. Buldyrev, S. Havlin, H.E. Stanley, Phys. Rev. Lett. (1997), in press.
[8] G.I. Bell, Comput. Chem. 20 (1996) 41.