Distribution of Base Pair Repeats in Coding and Noncoding DNA Sequences

Nikolay V. Dokholyan,[1] Sergey V. Buldyrev,[1] Shlomo Havlin,[1,2] and H. Eugene Stanley[1]
[1] *Center for Polymer Studies, Physics Department, Boston University, Boston, Massachusetts 02215*
[2] *Gonda-Goldschmied Center and Department of Physics, Bar-Ilan University, Ramat Gan, 52900 Israel*
(Received 24 April 1997)

We analyze the histograms for the lengths of the 16 possible distinct repeats of identical dimers, known as dimeric tandem repeats, in DNA sequences. For *coding* regions, the probability of finding a repetitive sequence of ℓ copies of a particular dimer decreases exponentially as ℓ increases. For the *noncoding* regions, the distribution functions for most of the 16 dimers have long tails and can be approximated by power-law functions, while for *coding* DNA, they can be well fit by a first-order Markov process. We propose a model, based on known biophysical processes, which leads to the observed probability distribution functions for *noncoding* DNA. We argue that this difference in the shape of the distribution functions between coding and noncoding DNA arises from the fact that noncoding DNA is more tolerant to evolutionary mutational alterations than coding DNA. [S0031-9007(97)04907-7]

PACS numbers: 87.10.+e

Interest in the growth and evolution of simple sequence repeats in DNA sequences is increasing due to their important role in genetic diseases, genome organization, and evolutionary processes [1,2]. One intriguing property of simple repeats is that they constitute a large fraction of *noncoding* DNA, but are relatively rare in protein *coding* sequences [3]. Another reason for the interest in simple sequence repeats is their possible relation to the long-range correlations found in DNA sequences: recent studies [4,5] support the claim [6,7] that the range of correlations in nucleotide composition is longer in noncoding regions than in coding ones.

Here we study the length distribution functions of these simple repeats, and we propose a model of DNA evolution which leads to the observed distributions. Specifically, we consider the distribution of repeats of identical dimers, called dimeric tandem repeats (DTR). DTR are so abundant in noncoding DNA that their presence can be observed by global statistical methods such as the power spectrum [8], which reveals a peak at frequency 1/2 for noncoding DNA (corresponding to repetition of dimers) and the absence of this peak in *coding* DNA (see, e.g., Fig. 1 of [4]). This difference in the abundance of DTR in coding DNA and noncoding DNA suggests that these repeats may play a role in the organization and evolution of DNA.

We analyze all vertebrate, invertebrate, mammal, primate, and plant taxonomic partitions of the GenBank release 96.0 and construct the length histograms $N_{xy}(\ell)$ of the 16 possible DTR, where ℓ is the number of identical copies of a particular dinucleotide xy, and x, y are the letters A, C, G, and T of the DNA "alphabet"—e.g., $AT\,AT\,AT = (AT)_{\ell=3}$, $CCCC = (CC)_{\ell=2}$.

We find two principal results:

(i) Coding.—All 16 DTR in *coding* DNA have distribution functions not significantly different from those of an uncorrelated or short-range correlated random sequence [Fig. 1(a)]. Thus

$$N_{xy}(\ell) \sim \exp(-k^0_{xy}\ell)\,, \tag{1}$$

where k^0_{xy} is the logarithm of the concentration of dimer xy. The exponential distributions of DTR in protein coding sequences are consistent with the hypothesis of strong evolutionary pressure against DTR expansion in active proteins [9].

(ii) Noncoding.—The length distributions in *noncoding* DNA for most DTR decay much more slowly than exponentially [Figs. 1(b) and 2]. With the exception of three cases—CC, CG, and GG—the DTR length distribution functions can be better approximated by power-law functions

$$N_{xy}(\ell) \sim \ell^{-\mu}, \tag{2}$$

where μ ranges from 2 to 4.5 [10] depending on the taxonomic class and type of DTR. According to the theory of Lévy walks, in the case $2 < \mu < 3$, the power-law distribution of simple repeats leads to the existence of long-range power-law correlations [11]. We note that the abundance of long dimeric repeats in noncoding DNA contributes to the presence of long-range correlations, while the lack of long dimeric repeats in coding DNA is related to the absence of long-range correlations [12]. However, long-range correlations in noncoding DNA are not only due to exact repetitions of dimers; other types of repeats occur, including trimer repeats, nonperfect simple repeats (simple repeats with a few substitutions), transposable elements [13], and long runs of purines and pyrimidines [14].

Next we discuss these two results in detail.

(i) Coding.—The exponential distribution of DTR in *coding* DNA reflects the fact that DTR represent uncorrelated or short-range correlated sequences which

 0031-9007/97/79(25)/5182(4)$10.00

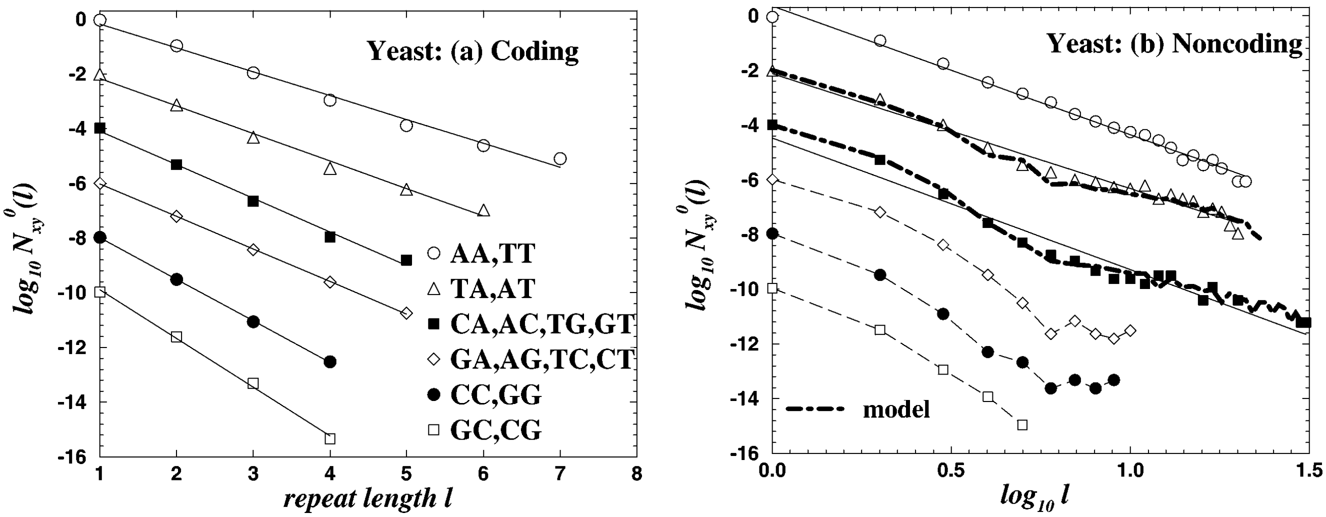

FIG. 1. The combined results for six groups of DTR: AA, TT (AA or TT) (○), TA, AT (△), CA, AC, TG, GT (■), CC, GG (●), GA, AG, TC, CT (◇), and GC, CG (□) for the complete yeast genome (total length 6 325 440 base pairs (bp); total length of the *coding* DNA 3 549 593 bp). We use this classification because A is complementary to T, and C is complementary to G; also, we average over two possible directions of reading of DNA sequences. For clarity, we separate plots for these six groups by shifting them by $\log_{10}(100) = 2$. (a) Semilogarithmic plot of $N^0_{xy}(\ell) \equiv N_{xy}(\ell)/N_{xy}(1)$ for coding DNA. Note that the data fall on straight lines, so Eq. (1) holds. In all 16 yeast chromosomes there are ten occurrences of DTR of length greater or equal to 10, which are not shown on the graph because they belong to putative genes (they are denoted by "*not experimental*" in the GenBank database). (b) Double logarithmic plot of $N^0_{xy}(\ell)$ for the *noncoding* DNA. Now, the data for the first three groups can be fit by a straight line, so Eq. (2) approximately holds. We find that the slope $\mu = 4.7$, 4.2, and 4.8 for the first three groups of DTR [25]. The other three groups of DTR cannot be fit by power-law functions [26]. As an example of $\pi(r, \ell)$ being a function of both r and ℓ, we include the results of simulations (dot-dashed bold line) fitting the second and the third groups of repeats. For TA, AT repeats, $\pi(r, \ell)$ depends on ℓ as a step function: for $1 < r \leq 2$: $\pi(r, \ell) = 0.15$, when $0 < \ell < 6$; 0.60, when $6 \leq \ell < 13$; and 0, when $\ell \geq 13$. For $0 < r \leq 1$: $\pi(r, \ell) = 0.85$, when $0 < \ell < 6$; 0.40, when $6 \leq \ell < 13$; and 1, when $\ell \geq 13$. For CA, AC, TG, GT repeats, $\pi(r, \ell)$ is also a step function of ℓ: for $1 < r \leq 2$: $\pi(r, \ell) = 0.016$, when $0 < \ell < 5$; 0.48, when $5 \leq \ell < 8$; 0.32, when $8 \leq \ell < 18$; and 0, when $\ell \geq 18$. For $0 < r \leq 1$: $\pi(r, \ell) = 0.984$, when $0 < \ell < 5$; 0.52, when $5 \leq \ell < 8$; 0.68, when $8 \leq \ell < 18$, and 1, when $\ell \geq 18$. In case of both groups of repeats, we start from a random sequence with equal concentration of all dimers $1/16 = 0.0625$ and produce 10^6 iterations of the random multiplicative process.

can be described by a Markov process [15], defined by a 16×16 matrix Π, whose elements $\Pi_{(xy)(zw)}$ are the conditional probabilities of finding a dimer zw after a dimer xy. The length distribution function of dimeric tandem repeats $(xy)_\ell$ of length ℓ is

$$N(\ell) = L p_{xy} \Pi^{\ell-1}_{(xy)(xy)} (1 - \Pi_{(xy)(xy)})^2 \sim 10^{-|k^M_{xy}| \cdot \ell}, \quad (3)$$

where $k^M_{xy} \equiv \log_{10} \Pi_{(xy)(xy)}$, L is the length of DNA sequence, and p_{xy} is the probability of finding a dimer xy in the large L limit [16]. In cases where the semilogarithmic plot of $N(\ell)$ is a straight line, we find that the actual slopes k_{xy} (which we calculate by linear regression) does not differ from k^M_{xy} by more than 10%, i.e., $|k_{xy} - k^M_{xy}|/k_{xy} < 0.1$ [17].

(ii) Noncoding.—Previous models of simple sequence repeat expansion do not resolve the question of long tails of DTR length distribution functions in noncoding DNA [18]. Here, we develop a model that reproduces the observed DTR distributions. We assume that in a single mutation, a repeat of length ℓ can expand or contract to a repeat of length $r\ell$, with conditional probability $\pi(r, \ell)$, where

$$\int_0^\infty \pi(r, \ell)\, dr = 1 \, . \quad (4)$$

The growth ($r > 1$) or contraction ($r < 1$) of the repeat can be caused by several types of mutations, such as unequal chromosomal crossing over [13,18] when a parental chromosome can elongate or shrink by some fraction of its length, or slippage during replication (see [19,20], and references therein), which leads to insertions or deletions of large fractions of repeats.

After t steps of evolution the length of the repeat is given by $\ell = \ell_0 \cdot \prod_{i=1}^{i=t} r_i$. Such a random multiplicative process leads, in many cases, to a stable distribution of repeat length $N(\ell)$ in the long time limit ($t \to \infty$). To avoid extinction of repeats, one can either (a) set a nonzero probability to reappear, or (b) set $\pi(r, \ell) = 0$ when $rl \leq 1$ [21].

It is impossible to find $N(\ell)$ analytically in the general case. However, for the case where this conditional probability $\pi(r, \ell)$ is a function only of r, i.e., can be written in the form

$$\pi(r, \ell) = \begin{cases} C(r), & r\ell > 1 \, , \\ 0, & r\ell \leq 1 \, , \end{cases} \quad (5)$$

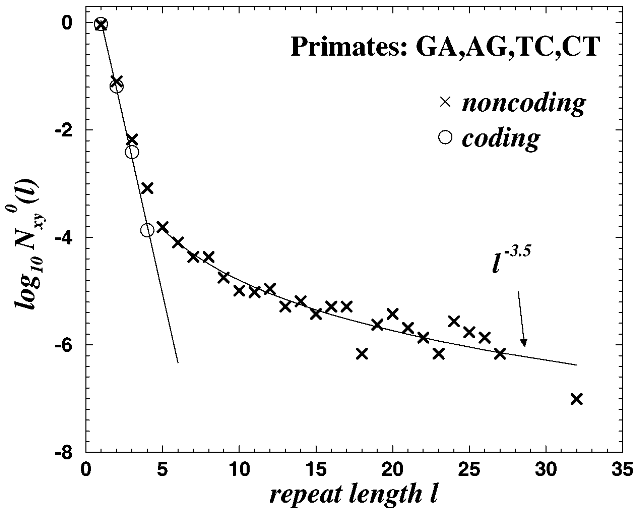

FIG. 2. Semilogarithmic plot of $N_{xy}^0(\ell)$ for all primate sequences in GenBank; as an example we show the combined results for GA, AG, TC, CT. Note that the data for coding DNA fall on a straight line, so Eq. (1) holds. The data for noncoding DNA can be approximately fit by a power-law function, so Eq. (2) holds. We find that the exponent $\mu = 3.5$ for fitting range $\ell \in [5, 35]$, with confidence value $R = 0.98$. The exponent is sensitive to the fitting range, e.g., if we concentrate our attention at the tail of the distribution $\ell \in [10, 35]$ we get exponent $\mu = 3$. The difference of length distribution functions for coding and noncoding DNA is dramatic; one can observe a DTR of length of ten dimers in noncoding DNA (with probability, roughly, $p \approx 10^{-5}$), while it is 7 orders of magnitude less probable to find such DTR in coding DNA ($p \approx 10^{-12}$).

the dynamics in terms of a new variable $z \equiv \ln \ell$ becomes a random additive process—i.e., simple diffusion in a semi-infinite space $z \geq 0$ with a reflecting wall at $z = 0$ [22] and an attractive uniform potential [23]. The length distribution function $\bar{N}(z)$ for the new variable z is given by

$$\bar{N}(z) \sim e^{-kz}, \tag{6}$$

where k is a constant which depends on the conditional probability $C(r)$ [24]. Since $z = \ln \ell$, Eq. (6) can be rewritten in the power-law form,

$$N(\ell) \sim \ell^{-\mu}, \tag{7}$$

which agrees with the experimental findings, Eq. (2). Here $\mu = k + 1$, because $N(\ell)d\ell = \bar{N}(z)dz$, and $dz = d\ell/\ell$.

Next we relate the value of μ to the specific form of $C(r)$. To this end, we introduce the time dependent distribution $\bar{N}(z,t)$. From the master equation for the z variable, written in the continuum limit,

$$\frac{d\bar{N}(z,t)}{dt} = -\bar{N}(z,t) + \int_{-\infty}^{+\infty} \bar{N}(z-x,t)C(x)\,dx\,, \tag{8}$$

and the stationarity condition

$$d\bar{N}(z,t)/dt = 0\,, \tag{9}$$

it follows that in the case where $\pi(r,\ell)$ is of the form of Eq. (5), $N(\ell)$ exhibits a power-law behavior with μ being a root of the equation

$$\int_0^\infty r^{\mu-1}C(r)\,dr = 1\,. \tag{10}$$

Equation (10) has a trivial solution $\mu = 1$, due to the normalization (4). Depending on the function $C(r)$, there may be, in addition, $\mu > 1$ roots of Eq. (10).

To illustrate, suppose

$$C(r) = \pi_1\delta(r - 1/2) + \pi_2\delta(r - 2)\,, \tag{11}$$

where $\pi_1 + \pi_2 = 1$ and $\delta(r)$ is the Dirac delta function. Then, Eq. (10) can be written as

$$\pi_1(1/2)^{\mu-1} + \pi_2 2^{\mu-1} = 1\,, \tag{12}$$

which has a root $\mu = 1 + \log_2(\pi_1/\pi_2)$. Note that μ ranges from 2 to 4.5 if π_1/π_2 ranges from 2 to 11.3. The fact that π_1 is greater than π_2 means that the probability for a repeat to shrink is larger than the probability to expand, which is biologically plausible since the repeats are preserved from unlimited expansion—i.e., the average repeat length does not diverge. In mathematical terms, the restriction $\pi_1/\pi_2 > 1$ is a necessary condition to obtain a stable probability distribution function, for otherwise condition (9) is not satisfied. This example shows that the model can produce power-law distributed repeats with any given exponent $\mu > 1$. The explanation of the empirical distributions for various kinds of repeats requires further study, which should take into account their specific biophysical and biochemical properties.

The assumption that $\pi(r,\ell) \equiv C(r)$ is independent of ℓ is an approximation. Because of the specific biochemical mechanisms, mutation rates may depend on the length ℓ of the DTR [20]. Even though this ℓ dependence causes deviation of the model distributions from a power-law behavior, the power-law functions often provide satisfactory approximations to the experimental data [25].

To further test the model, we compute numerically the DTR length distributions in the case when $\pi(r,\ell)$ depends both on r and ℓ. As it follows from our analysis, the length distribution of various DTR differ significantly from each other in various organisms. Varying $\pi(r,\ell)$, we find even better agreement with experimental data [Fig. 1(b)] for each particular case. The nonuniversality of $\pi(r,\ell)$ is biologically plausible since the mutation rates may strongly depend on the organism as well as on the repeat type. However, such mutation rates are presently not known. Therefore, the model can serve as a tool for determining and testing the mutation rates by fitting the experimental data.

In summary, we propose a mathematical description of actual mutational processes for noncoding DNA and show that these processes can produce nonexponential, broad tails of DTR length distribution functions. In contrast, coding DNA are more preserved from the DTR expansion since such mutations would lead to a nonfunctional protein and as a result to the extinction of the organism. We argue that this fact explains the exponential DTR distribution functions in coding sequences. The properties of DTR may serve as additional information in various applications such as distinguishing between coding and noncoding DNA and understanding the molecular evolution.

We thank N. Broude, R. S. Dokholyan, I. Große, P. L. Krapivsky, C.-K. Peng, G. M. Viswanathan, and R. Wells for fruitful discussions, and G. H. Weiss for seminal contributions in the formative stages of this work. This work is supported by NIH-HGP.

---

[1] Y. Ionov, M. A. Peinado, S. Malkhosyan, D. Shibata, and M. Perucho, Nature (London) **363**, 558 (1993).

[2] A. M. Bowcock *et al.,* Nature (London) **368**, 455 (1994).

[3] J. Jurka and C. Pethiyagoda, J. Mol. Evol. **40**, 120 (1995). See also the earlier work of K. A. Marx, S. T. Hess, and R. D. Blake, J. Biomol. Struct. Dyn. **11**, 57 (1993).

[4] S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, M. E. Matsa, C.-K. Peng, M. Simons, and H. E. Stanley, Phys. Rev. E **51**, 5084 (1995); S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, and H. E. Stanley, in *Fractals in Science,* edited by A. Bunde and S. Havlin (Springer-Verlag, Berlin, 1994).

[5] A. Arneodo, E. Bacry, P. V. Graves, and J. F. Muzy, Phys. Rev. Lett. **74**, 3293 (1995).

[6] C. K. Peng, S. Buldyrev, A. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, Nature (London) **356**, 168 (1992).

[7] W. Li and K. Kaneko, Europhys. Lett. **17**, 655 (1992).

[8] For coding DNA the power spectrum has a pronounced peak at $1/3$ bp^{-1} (corresponding to 3-tuples or trimers), suggesting the presence of codon structure that can be used as a strong indicator of coding DNA: J. W. Fickett and C. S. Tung, Nucleic Acids Res. **20**, 6441 (1992).

[9] E. I. Shakhnovich and A. M. Gutin, Nature (London) **346**, 773 (1990).

[10] For example, for TA, AT DTR in noncoding vertebrate DNA the length distribution exhibits power-law tail with the exponent $\mu = 2.7$.

[11] J. Klafter, G. Zumofen, and M. F. Shlesinger, in *Lévy Flights and Related Topics in Physics,* edited by M. F. Shlesinger, G. M. Zaslavsky, and U. Frisch (Springer-Verlag, Berlin, 1995).

[12] G. M. Viswanathan, S. V. Buldyrev, S. Havlin, and H. E. Stanley, Biophys. J. **72**, 866 (1997).

[13] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson, *Molecular Biology of the Cell* (Garland, New York, 1994).

[14] G. Yagil, Yeast **10**, 603 (1994).

[15] B. V. Gnedenko, *The Theory of Probability* (Chelsea, New York, 1968).

[16] This probability p_{xy} is an eigenvector of the matrix Π corresponding to the largest eigenvalue $\lambda = 1$ (see [15]).

[17] The only exception is the repeat AA for which $|k - k_M|/k \approx 0.2$. Note that since in the case of uncorrelated random sequence $\Pi_{(xy)(xy)} = p_{xy}$, k_M for a certain DTR may coincide with logarithm of concentration of the corresponding dimer.

[18] G. I. Bell and J. Jurka, J. Mol. Evol. **44**, 414 (1997), and references therein.

[19] R. I. Richards and G. R. Sutherland, Nature Genetics (London) **6** 114 (1994).

[20] R. D. Wells, J. Biol. Chem. **271**, 2875 (1996).

[21] Both ways are mathematically equivalent, since we can always control the probability of creation of new repeats at $\ell = 1$ using condition (a) so that we reach condition (b). These conditions, (a) or (b), might be biologically caused by point mutations [13]—the random substitution of a nucleotide by other ones, since they can create repeats of length $\ell = 1$.

[22] $\ell \geq 1$ corresponds to $z \geq 0$, which reflects the fact that the DTR cannot disappear.

[23] A classical example of such a process is Brownian motion in a constant potential field, which leads in the continuum limit to an exponential decrease of the atmosphere's density.

[24] Note that reflecting boundary condition is crucial for this model. In the case when we do not have boundaries, diffusion in the presence of the constant field yields $\bar{N}(z,t) = (\sqrt{2\pi DT})^{-1} \exp[-(z - vt)^2/2Dt]$, where D and v are parameters of the field. The reflecting boundary condition is a special condition, which produces the exponential solution (6).

[25] The correlation coefficients for the first three groups of DTR are $R = 0.99$, 0.99, and 0.98. For comparison, the fit of DTR length distribution functions to an exponential (on semilogarithmic plot) is not nearly as good, yielding $R = 0.93$, 0.90, and 0.80 for these groups of DTR. The fitting range is $\ell \in [1, 31]$ for both methods of fitting.

[26] Note that for the *noncoding* yeast DNA we find that the distributions of GA, AG, TC, CT do not have power-law tails, while for vertebrates (see Fig. 2) they do.