ELSEVIER

Physica A 273 (1999) 19–32

PHYSICA A

www.elsevier.com/locate/physa

Expansion of tandem repeats and oligomer clustering in coding and noncoding DNA sequences

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

[a] *Center for Polymer Studies and Department of Physics Boston University, Boston, MA 02215, USA*
[b] *Gonda-Goldschmied Center and Department of Physics Bar-Ilan University, Ramat Gan, Israel*
[c] *Chemistry Department, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*

Received 10 August 1999

Abstract

We review recent studies of distribution of dimeric tandem repeats and short oligomer clustering in DNA sequences. We find that distribution of dimeric tandem repeats in coding DNA is exponential, while in noncoding DNA it often has long power-law tails. We explain this phenomenon using mutation models based on random multiplicative processes. We also develop a clustering measure based on percolation theory that quantifies the degree of clustering of short oligomers. We find that mono-, di-, and tetramers cluster more in noncoding DNA than in coding DNA. However trimers have some degree of clustering in coding DNA and noncoding DNA. We relate this phenomena to modes of tandem repeat expansion. © 1999 Elsevier Science B.V. All rights reserved.

PACS: 87.14.G; 87.23; 64.60.A

Keywords: DNA; Evolution; Oligomer clustering; Percolation theory

1. Introduction

The origin, evolution, and biological role of tandem repeats in DNA, also known as microsatellites or simple sequence repeats (SSR), are presently one of the intriguing puzzles of molecular biology. The expansion of such SSR has recently become of great

[*] Corresponding author. Fax: +1-617-353-9393.
E-mail address: sergey@miranda.bu.edu (S.V. Buldyrev)

PII: S0378-4371(99)00338-6

interest due to their role in genome organization and evolutionary processes [1–11]. It is known that SSR constitute a large fraction of noncoding DNA and are relatively rare in protein coding sequences.

SSR are of considerable practical and theoretical interest due to their high polymorphism [7]. The formation of a hairpin structure during replication [12,13] is believed to be the cause of the *CNG* repeat expansions, which are associated with a broad variety of genetic diseases. Dimeric SSR of the type $(CA)_\ell$ are also known to expand due to slippage in the replication process. These errors are usually eliminated by the mismatch-repair enzyme MSH2. However, a mutation in the MSH2 gene leads to an uncontrolled expansion of repeats – a common cause of ovarian cancers [14]. Similar mechanisms are attributable for other types of cancer [6,15–17]. Telomeric SSR, which control DNA sequence size during replication, illustrate another crucial role of tandem repeats [18].

A study of SSR from primates, emphasizing their abundance, length polymorphism, and overall tendency to expand in different sequence contexts, was reported by Jurka and Pethiyagoda [10]. The probability distribution functions for the length of special classes of repeats have been studied in many publications (see, e.g., Refs. [19–22]. Bell and Jurka [23] studied the length distributions of dimeric tandem repeats of rodent and primate DNA. Other studies, reporting periodicities of various SSR in introns, have been carried out [24,25]. An analysis of clustering of nucleotides has been done by Mrazek and Kypr [26] and by Lio et al. for *Haemophilus influenzae* and *Saccharomyces cerevisiae* chromosomes [27]. Systematic analysis of SSR distributions has yet to be done and is the focus of this paper. Specifically, we consider the distribution of the most simple case of SSR – repeats of identical dimers, i.e., dimeric tandem repeats (DTR). It is possible to construct a statistically significant length distribution of SSR only in the cases of mono- and di-nucleotides.

In order to extend the study of patterns of nucleotides in DNA, we develop a quantitative method for studying the repetitions of oligomers (mono-, di-, tri-, and tetranucleotides) in coding and noncoding DNA. Using the concepts of percolation theory [28–30], we calculate the mean length (defined below) of repetitions of oligomers. We also calculate the expected length of repetitions if the oligomers – with the same frequencies as in real sequence – were randomly placed along an artificial sequence. We use the expected length of repetitions of oligomers as a control. By forming the dimensionless ratio between the actual value to the control value, we can recognize whether oligomers “cluster” (repeat more than they would if their order were randomly shuffled) or “repel” (repeat less than they would if their order were randomly shuffled). In such a way we can understand if oligomers in DNA tend to aggregate or segregate.

Section 2 describes methods we use to analyze distributions of DTR. The difference between coding and noncoding DNA sequences is discussed in Section 3. The relation to the evolutionary mechanisms of DTR expansion is discussed in Section 4. The methods and results for clustering measures are presented in Sections 5 and 6.

2. DTR methods

We study [31] genomes of four different organisms: human (*homo sapiens*), mouse (*mus musculus*), nematode (*caenorhabditis elegance*) and yeast (*saccharomyces cerevisiae*). In order to minimize the artificial statistical bias of the GenBank towards specific proteins, we restrict our study to the complete genomes of yeast (all 16 chromosomes) and C. elegans (all six chromosomes), and large human and mouse genomic sequences exceeding 100 kbp in length. We found 1113 such sequences of total length 144,833,600 bp in human and 33 such sequences of total length 4,545,649 bp in mouse DNA sequences (GenBank release 110.0).

We analyze separately all coding and noncoding regions (intergenic and introns) for each of the organisms. We identify coding DNA using CDS key word in the GenBank flat file format. We concatenate sections that belong to one CDS and correspond to the genetic code for one protein. We identify introns as sections between exons within one CDS. The region between different CDS are identified as intergenic. Thus, the intergenic regions, by our definition, also include unrestricted $5'$ and $3'$ ends of the genes.

First, we calculate number of occurrences $N(\ell)$ of dimeric tandem repeats of ℓ repetitions for 16 types of dimers. We combine results for six groups of DTR: (1) AA, TT (AA or TT); (2) TA, AT; (3) CA, AC, TG, GT; (4) CC, GG; (5) GA, AG, TC, CT; and (6) GC, CG. We use this classification because A is complementary to T, and C is complementary to G; and, we average over two possible directions of reading DNA sequences. In addition, we combine data for repeats xy and yx, where x and y denote nucleotides A, C, G or T, since repeats xy and yx have almost identical distributions. In fact, repeat $(xy)_\ell$ must become $(yx)_{\ell\pm1}$ if one shifts the reading frame by one bp.

Next, we calculate the normalized number of repeats $N_0(\ell) = N(\ell)/N(1)$ of length ℓ, where $N(1)$ is the total number of occurrences of a single dimer. If there are no repeats for one or more consecutive values of ℓ between points ℓ' and ℓ'', we substitute $N_0(\ell'')$ by $N_0(\ell'')/(\ell'' - \ell')$.

3. DTR results

We find that the normalized number of repeats $N_0(\ell)$ for all six groups of dimeric tandem repeats in coding DNA in all four analyzed organisms decays rapidly with ℓ. By plotting $N_0(\ell)$ in the semilogarithmic scale (Figs. 1a and 2a) we observe that all of these functions have a linear decay, indicating exponential functional form of $N_0(\ell) \sim \exp(-k\ell)$, which is in agreement with predictions of short-range markovian models. In mouse and human we find practically no deviations from exponential behavior for coding regions. For yeast and C. elegans we find only 13 and 19 CDS regions correspondingly which have repeats of length larger than 10.

For noncoding regions, distributions $N_0(\ell)$ can be better described by a power law:

$$N_0(\ell) \sim \ell^{-\mu} \tag{1}$$

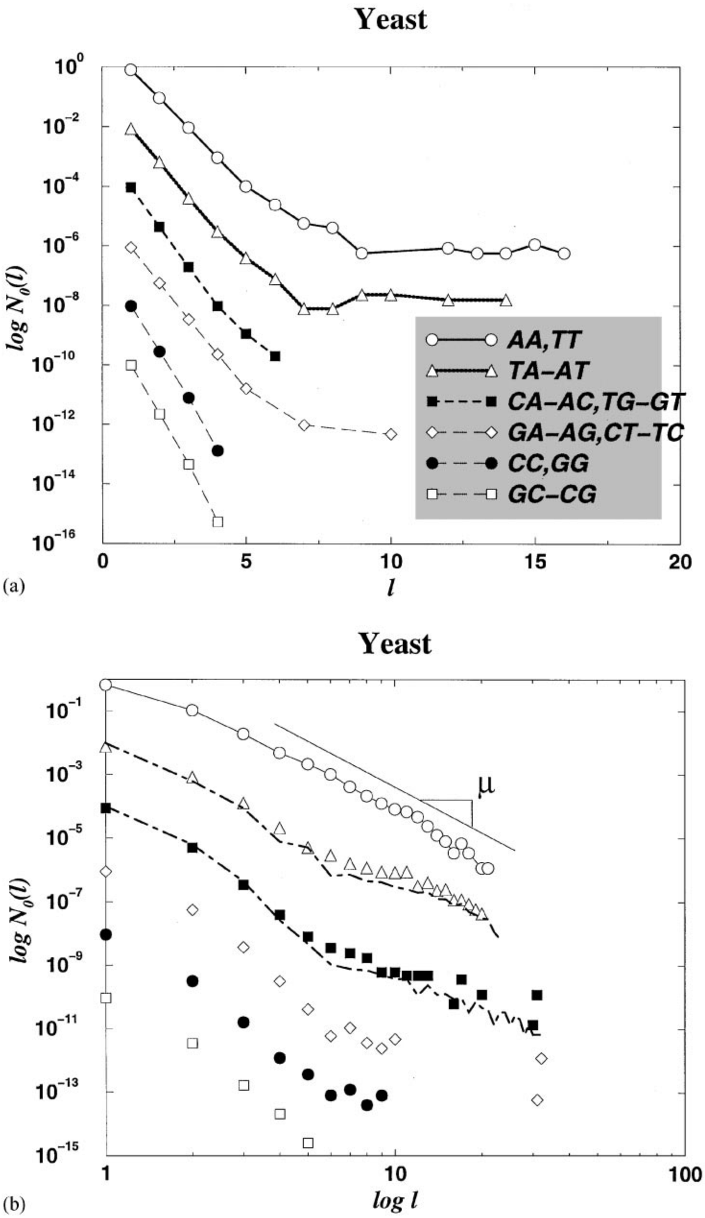

Fig. 1. The combined plot of average normalized number of repeats for six groups of dimeric tandem repeats for the complete yeast genome: AA, TT (AA or TT) (○); TA, AT(△); CA, AC, TG, GT (■); CC, GG (●); GA, AG, TC, CT (◇); and GC, CG (□) (a) for coding DNA in semilogarithmic scale and (b) for noncoding DNA in double-logarithmic scale. For clarity, we separate plots for these six groups by shifting them by a factor of 100 on the ordinate. The straight line on plot (b) indicates a power-law function $f(x) \sim x^{-\mu}$ with $\mu = 5.3$. In (b) as an example of $P(r, \ell)$ 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 (see Fig. 3). The values of μ for first three groups of repeats are 5.3, 3.2, 2.8, from top to bottom, fitting range is $\ell > 5$.

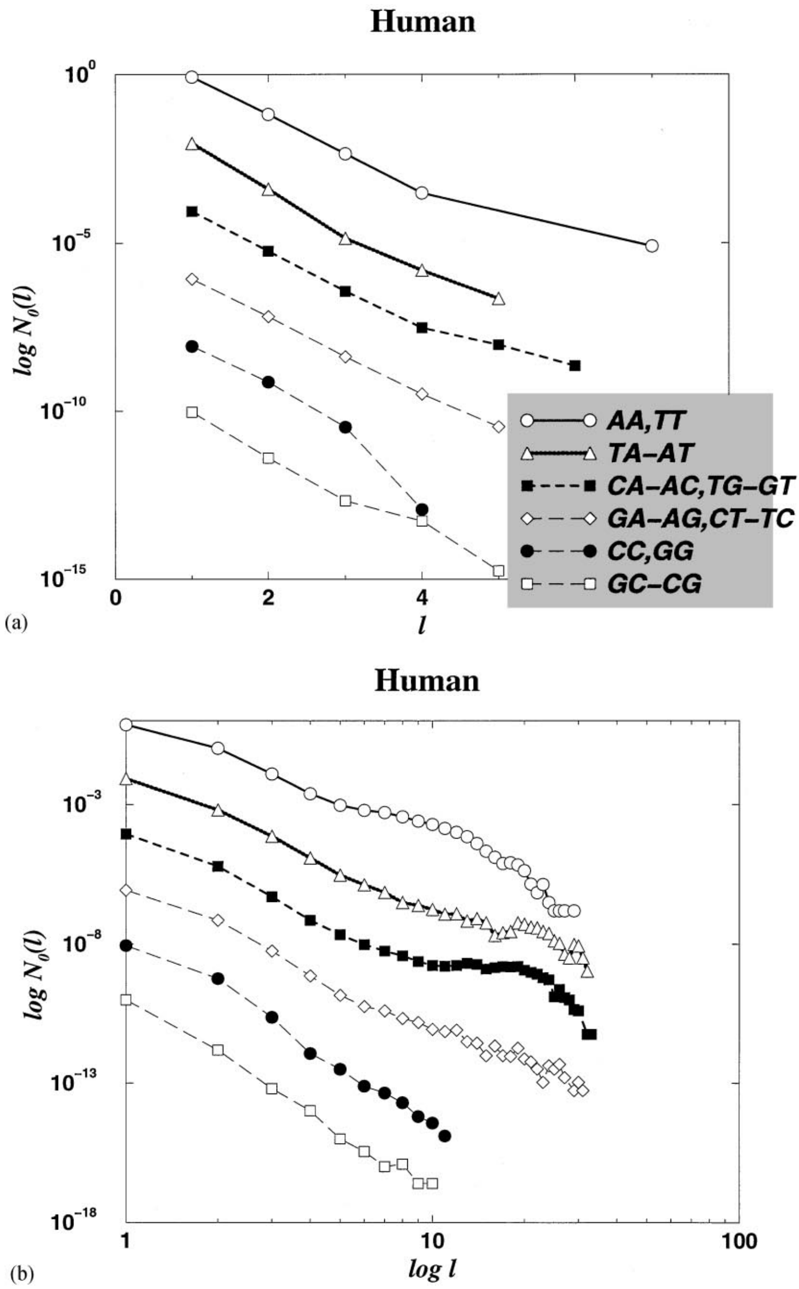

Fig. 2. The same as above but for known human genome. The values of μ for these six groups of repeats are 5.6, 3.3, 3.2, 4.1, 6.7, 5.4 from top to bottom, fitting range is $\ell > 5$. (a) Coding; (b) non-coding.

with μ ranging between 2 and 5 for different organisms. In this case, these distributions should be straight lines on a double logarithmic plot. In fact, that is what we observe from Figs. 1b and 2b. We do not observe a significant statistical difference in the functional behavior of $N_0(\ell)$ for introns and intergenic sequences, so we present combined data for noncoding DNA (introns and intergenic sequences).

There are two exceptions:

(i) We find that in some cases, the double logarithmic plots for noncoding regions have long tails which are almost horizontal straight lines with a rapid cut-off at about 30 copies, and cannot be approximated by a power-law function. These plateaus correspond to noncoding sequences of the dimers *CA–AC*, *TG–GT*, and *GA–AG*, *TC–CT* in mouse and human DNA (see Fig. 2b). This effect is not observed in C. elegance and yeast.

(ii) We also observe that all organisms lack *CG–GC* repeats in the intron and noncoding DNA.

4. DTR model

Recently, several mechanisms of simple sequence repeat expansion have been proposed [2,5,6,9,13,21,22,32–34]. The model, proposed recently to explain power-law distributed repeats [21] can produce power-law distributed repeats with any given exponent μ [for details see Ref. [22]].

The mechanism proposed in Refs. [21,22] is based on random multiplicative processes, which can reproduce the observed non-exponential distribution of repeats. The increase or decrease of repeat length can occur due to unequal crossover [18,35] or slippage during replication [5,13,36].

It is reasonable to assume (see Wells, 1996, and references therein) that in these types of mutations, the new length ℓ' of the repeat is not a stepwise increase or decrease of the old length, but is defined as a product $\ell' = \ell r$, where r is some random variable. This variable r is distributed according to the probability distribution function $P(r,\ell)$, which depends both on ℓ and r. As has been shown in Refs. [21,22,37], in the most simple case when $P(r,\ell)$ does not depend on ℓ (we denote it by $P(r)$), the model produces pure power-law distribution of repeats, and the value of a power-law exponent μ can be determined from the equation

$$1 = \int_0^{+\infty} P(r) r^{\mu-1}\,\mathrm{d}r\,. \tag{2}$$

The difference between the empirical distributions for various kinds of repeats can be attributed to the fact that the probability rates $P(r,\ell)$ of various mutations strongly depend on the length of the repeats ℓ [21,22,36,13], i.e., there exist biochemical dependence on the repeat size. For example, the power-law behavior usually starts from repeats of length $\ell \geqslant 5$. This may indicate that short repeats are distributed exponentially, as in random sequence. Thus, it is plausible to conclude that the mutation processes target repeats of length above $\ell \geqslant 5$.

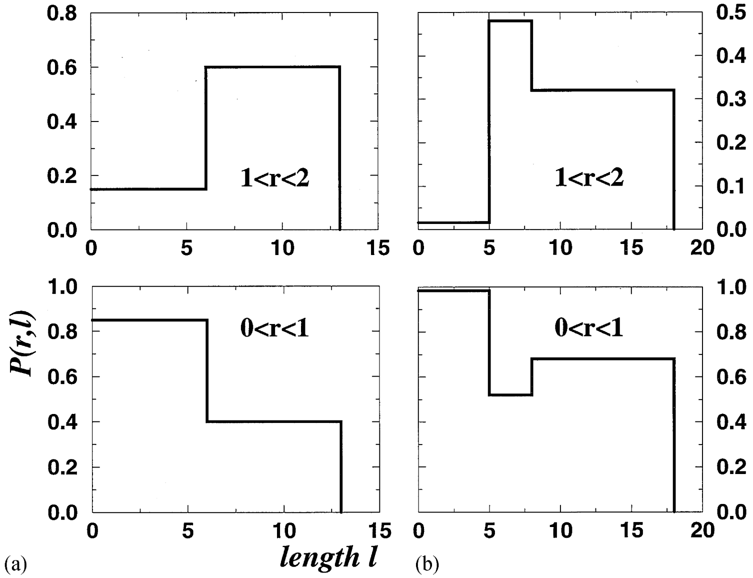

Fig. 3. As an example of $P(r,\ell)$ being a function of both r and ℓ, we use following $P(r,\ell)$ to fit TA, AT, CA, AC, TG, GT repeats in noncoding yeast DNA (see Fig. 1b). (a) For TA, AT repeats, $P(r,\ell)$ depends on ℓ as a step function: for $1 < r \leqslant 2$: $P(r,\ell) = 0.15$, when $0 < \ell < 6$; 0.60, when $6 \leqslant \ell < 13$; and 0, when $\ell \geqslant 13$. For $0 < r \leqslant 1$: $P(r,\ell) = 0.85$, when $0 < \ell < 6$; 0.40, when $6 \leqslant \ell < 13$; and 1, when $\ell \geqslant 13$. For $r > 2$ we assume $P(r,\ell) = 0$, which means that a repeat cannot expand by more than a factor of 2 in a single mutation. (b) For CA, AC, TG, GT repeats, $P(r,\ell)$ is also a step function of ℓ: for $1 < r \leqslant 2$: $P(r,\ell) = 0.016$, when $0 < \ell < 5$; 0.48, when $5 \leqslant \ell < 8$; 0.32, when $8 \leqslant \ell < 18$; and 0, when $\ell \geqslant 18$. For $0 < r \leqslant 1$: $P(r,\ell) = 0.984$, when $0 < \ell < 5$; 0.52, when $5 \leqslant \ell < 8$; 0.68, when $8 \leqslant \ell < 18$; and 1, when $\ell \geqslant 18$. In case of both groups of repeats, we start from a random sequence with equal concentration of all dimers $\frac{1}{16} = 0.0625$ and produce 10^6 iterations of the random multiplicative process.

It was shown by computer simulations [21] that imposing the length constraints on the mutational rates $P(r,\ell)$ produces better fit of experimental data (see Fig. 1b. For example, the probability distribution function $P(r,\ell)$, has been chosen as shown in Fig. 3 to fit distributions of TA, AT and CA, AC, TG, GT repeats in yeast (Fig. 1b). The rigorous modeling of the specific tandem repeats still requires further investigation taking into account their particular biophysical and biochemical properties.

A different model was proposed by [34], which was also able to reproduce long tails in the repeat length distribution. This model assumes the stepwise change in repeat length with the mutation rate proportional to the repeat length. It is possible to show that this model can be mapped to a random multiplicative process with a specific form of distribution $P(r,\ell)$, where $r = \ell'/\ell$, ℓ is the original length and ℓ' is a repeat length after a time interval during which several stepwise mutations can occur.

The advantage of the model proposed in [21] is that if we tune the dependence of mutation rates $P(r,\ell)$ on ℓ, we can precisely fit the distributions of the dimeric repeats (see Figs. 1b and 3). The "plateau" in the distributions of dimeric repeats, discussed

Table 1
The total length in bp of the coding and noncoding regions studied in clustering analysis. The protein coding sequences are constructed by concatenating sequences belonging to the same gene, denoted as *CDS* in the GenBank. The noncoding sequences are constructed by concatenating sequences, which are not denoted as *CDS* in the GenBank

Organism	Noncoding	Coding
Vertebrates	206,757	91,323
Primates	5,161,953	692,634
Invertebrates	5,322,555	6,993,572
Plants	738,506	4,946,293
Mammals	121,556	56,994
Rodents	1,131,628	253,200

in the Section 3, can also be explained by the mutational rate variability and can be fit by the model.

5. Cluster measures

We quantify [38] the repetitions of oligomers by dividing the sequence into the non-overlapping windows of n nucleotides, where n is the size of an oligomer. For trimers ($n=3$) we select biological reading frames when we study coding regions. In all other cases we select randomly chosen reading frame.

We analyze separately protein coding and noncoding sequences. For coding sequences we concatenate exons within a single gene (excluding the untranslated $5'$ and $3'$ ends). Noncoding sequences we identify as those that are not explicitly specified as CDS in the GenBank flat file format. In order to deal with the bias in the GenBank database due to the multiple entries of short copies of some fragments of the larger DNA sequences, we select only those loci that exceed in length 10^4 bp. This reduces the redundancy of the data we analyzed. The total length, L, and the number of sequences analyzed in coding and noncoding regions of different taxonomic partitions is reported in Table 1.

First, we compute the number of repeats of length ℓ of a given repeat in analyzed set of sequences: $N_i(\ell)$, where $i=1,\dots,M$ is the index of an oligomer and $M=4^n$ is the total number of distinct oligomers of size n. According to our definition, we have

$$\sum_{\ell=1}^{\infty}\sum_{i=1}^{M} N_i(\ell)\ell = L\,. \tag{3}$$

Next, we introduce two measures of repeat length: (i) We define the “number” average

$$\langle \ell \rangle_n \equiv \frac{L}{N}\,, \tag{4}$$

where

$$N=\sum_{\ell=1}^{\infty}\sum_{i=1}^{M}N_i(\ell) \tag{5}$$

is the total number of repeat occurrences.

(ii) We also define the "weight" average (see, e.g. [30]):

$$\langle\ell\rangle_w\equiv\frac{\sum_{\ell=1}^{\infty}\sum_{i=1}^{M}\ell^2N_i(\ell)}{\sum_{\ell=1}^{\infty}\sum_{i=1}^{M}\ell N_i(\ell)}=\frac{\sum_{\ell=1}^{\infty}\sum_{i=1}^{M}\ell^2N_i(\ell)}{L}\,. \tag{6}$$

This definition gives larger weights to longer repeats. The utility of Eq. (6) is that $\langle\ell\rangle_w$ is the average length of a repeat to which a randomly chosen oligomer belongs.

Next, we calculate the control value, where the control is obtained by scrambling (random reshuffling) the order of the oligomers. If all the nucleotides were evenly represented, each oligomer would have a frequency of $1/4^n$, where n is the size of the oligomer, e.g., $n=1$ for monomers, $n=2$ for dimers, etc. Since the frequencies of the nucleotides vary, we calculate the actual frequency of each oligomer in a particular set of DNA sequences. We generate a control sequence by random concatenation of the oligomers with given frequencies.

For the control sequence, the probability P_ℓ that a given oligomer belongs to a "cluster" (aggregate) of *exactly* ℓ repetitions in an uncorrelated random sequence is given by percolation theory [28,29]:

$$P_i(\ell)=\ell p_i^\ell(1-p_i)^2\,, \tag{7}$$

where p_i is the frequency of a particular oligomer. By multiplying $P_i(\ell)$ by the total number of oligomers in the set of sequences of size L, we find the total number of clusters of size ℓ. Thus, for random uncorrelated sequences, the expected number of clusters $N_i^0(\ell)$ of size ℓ (the control) is given by

$$N_i^0(\ell)=Lp_i^\ell(1-p_i)^2\,. \tag{8}$$

It is possible to calculate theoretical predictions for both measures of repeat lengths for a control sequence in which the order of oligomers is randomly scrambled. For such an uncorrelated random sequence we find

$$\langle\ell\rangle_n^{\rm th}=\frac{1}{1-\sum_{i=1}^{M}p_i^2}\,, \tag{9}$$

where p_i is the frequency of each oligomer, and

$$\langle\ell\rangle_w^{\rm th}=1+2\sum_{i=1}^{M}\frac{p_i^2}{1-p_i}\,. \tag{10}$$

To quantify the relative clustering strength, we introduce two *clustering ratios*, defined by

$$R_n\equiv\frac{\langle\ell\rangle_n-1}{\langle\ell\rangle_n^{\rm th}-1}\quad\text{and}\quad R_w\equiv\frac{\langle\ell\rangle_w-1}{\langle\ell\rangle_w^{\rm th}-1}\,. \tag{11}$$

The clustering ratios compare actual repeat length with the control case in which, by definition, no clustering occurs beyond the clustering that occurs in uncorrelated random process. Note, that for an uncorrelated random sequence the distributions of R_n and R_w are Gaussian, centered at $R_n = 1$ and $R_w = 1$ correspondingly, and standard deviations, inversely proportional to $\sqrt{L}$ (see Ref. [38]). In Table 2, we present the relative clustering ratios R_n and R_w.

6. Clustering results

We compare the ratios of the observed values of the average length $\langle l \rangle_n$ of oligomers (monomers, dimers, trimers, and tetramers) and their weight average $\langle l \rangle_w$ to the theoretically predicted for a randomly shuffled sequence. We consider primate, vertebrate, invertebrate, mammal, rodent, and plant taxonomic partitions of GenBank release 104. We limit our analysis only to eukaryotic genomes since for prokaryotic genomes our preliminary analysis shows virtually no clustering. The complete results for clustering ratio values and for the error bars of these values are presented in Table 2. To compute error bars we partitioned GenBank data sets into 10 subsets of size of 10% of the GenBank data sets. We compute the clustering ratios for each set and from the distribution of these values we determine the mean and the standard deviation, presented in Table 2. The probability that these distributions for coding and noncoding DNA belong to the same distribution is characterized by the p-value of the Kolmogorov–Smirnov test (see [39]).

(i) A significant difference between the clustering of monomers (excluding plants), dimers, and tetramers in coding versus noncoding DNA. The p-values for all the distributions of ratio value sets of above-mentioned groups of repeats do not exceed 2×10^{-5}.

(ii) The clustering ratios for the monomers in coding DNA for all the taxonomic partitions except plants are close to unity (within 9%), which means that they are close to being randomly distributed. For the noncoding DNA, however, these values are consistently greater than one, indicating the slight clustering of monomers.

(iii) The clustering ratios for the dimers in coding DNA are also close to unity (within 7%). However, these values are consistently smaller than unity, which indicates the slight repulsion of dimers in coding DNA. On the contrary, the clustering ratios for the dimers in noncoding DNA are consistently greater than unity. The clustering ratio values for the tetramers in coding DNA are consistently and significantly smaller than one (up to 32%) which indicates the repulsion of tetramers. On the contrary, the clustering ratios for the tetramers in noncoding DNA are consistently greater than unity.

(iv) The clustering ratios for the trimers for all organisms show strong clustering of the trimers in both coding and noncoding DNA. For primates and mammals, the Kolmogorov–Smirnov p-values for the R_n ratio are of the order of 1 (Table 2), which indicates that one cannot distinguish between coding and noncoding DNA based only on

Table 2
The average clustering ratio values R_n and R_w along with the error bars are shown for mono-, di-, tri-, and tetramers in coding and noncoding DNA of primate, vertebrate, invertebrate, mammal, rodent, and plant taxonomic partitions of the GenBank. The mean values and the error bars (one standard deviation) are computed by partitioning the GenBank data sets into 10 subsets of size 10% of the GenBank data sets, obtained for these independent subsets. Afterward, we compute R_n and R_w for each subset independently. Then we consider the distributions of the values of R_n and R_w for coding and noncoding DNA and compute the p-values for the Kolmogorov–Smirnov test indicating the probability that those R_n and R_w values (for coding and for noncoding DNA) are drawn from the same distribution. If p is close to 1, then the two distributions are drawn from the same distribution with the probability close to 1. If p is close to 0, then these distribution are taken from two different distributions with the probability $(1 - p) \approx 1$. These results are consistent with: (i) there is evolutionary pressure against clustering of repeats (except trimeric) in coding DNA; (ii) the clustering ratios for all organisms show strong clustering of the trimers; (iii) the difference between the clustering of trimers in coding DNA for different taxonomic partitions is less pronounced than in noncoding DNA

Organism	R_n			R_w		
	Coding	Noncoding	p-value	Coding	Noncoding	p-value
Monomers						
Primates	1.09 ± 0.01	1.26 ± 0.01	$< 2 \times 10^{-5}$	1.08 ± 0.01	1.43 ± 0.01	$< 2 \times 10^{-5}$
Vertebrates	1.03 ± 0.01	1.14 ± 0.01	$< 2 \times 10^{-5}$	1.02 ± 0.01	1.24 ± 0.01	$< 2 \times 10^{-5}$
Invertebrates	1.09 ± 0.01	1.40 ± 0.01	$< 2 \times 10^{-5}$	1.08 ± 0.01	1.58 ± 0.01	$< 2 \times 10^{-5}$
Mammals	1.06 ± 0.01	1.28 ± 0.02	$< 2 \times 10^{-5}$	1.04 ± 0.02	1.43 ± 0.04	$< 2 \times 10^{-5}$
Rodents	1.06 ± 0.01	1.18 ± 0.01	$< 2 \times 10^{-5}$	1.03 ± 0.01	1.30 ± 0.01	$< 2 \times 10^{-5}$
Plants	1.13 ± 0.01	1.10 ± 0.01	$< 2 \times 10^{-5}$	1.12 ± 0.01	1.17 ± 0.01	$< 2 \times 10^{-5}$
Dimers						
Primates	0.96 ± 0.01	1.39 ± 0.01	$< 2 \times 10^{-5}$	0.95 ± 0.01	1.73 ± 0.01	$< 2 \times 10^{-5}$
Vertebrates	1.00 ± 0.02	1.32 ± 0.02	$< 2 \times 10^{-5}$	1.00 ± 0.02	1.81 ± 0.13	$< 2 \times 10^{-5}$
Invertebrates	0.95 ± 0.01	1.39 ± 0.01	$< 2 \times 10^{-5}$	0.97 ± 0.01	1.49 ± 0.01	$< 2 \times 10^{-5}$
Mammals	0.94 ± 0.02	1.43 ± 0.04	$< 2 \times 10^{-5}$	0.94 ± 0.02	1.74 ± 0.09	$< 2 \times 10^{-5}$
Rodents	0.93 ± 0.01	1.47 ± 0.01	$< 2 \times 10^{-5}$	0.93 ± 0.01	2.33 ± 0.01	$< 2 \times 10^{-5}$
Plants	0.95 ± 0.01	1.21 ± 0.01	$< 2 \times 10^{-5}$	0.95 ± 0.01	1.36 ± 0.03	$< 2 \times 10^{-5}$
Trimers						
Primates	1.51 ± 0.02	1.49 ± 0.01	0.31	1.63 ± 0.03	1.91 ± 0.02	1.7×10^{-4}
Vertebrates	1.54 ± 0.05	1.40 ± 0.04	1.7×10^{-4}	1.68 ± 0.08	1.56 ± 0.06	3.1×10^{-2}
Invertebrates	1.49 ± 0.01	1.21 ± 0.01	1.7×10^{-4}	1.56 ± 0.01	1.27 ± 0.01	1.7×10^{-4}
Mammals	1.53 ± 0.04	1.51 ± 0.04	0.97	1.63 ± 0.06	1.87 ± 0.10	1.7×10^{-4}
Rodents	1.42 ± 0.02	1.40 ± 0.01	6.9×10^{-3}	1.52 ± 0.02	2.13 ± 0.07	$< 2 \times 10^{-5}$
Plants	1.42 ± 0.01	1.29 ± 0.02	1.7×10^{-4}	1.50 ± 0.01	1.45 ± 0.02	1.7×10^{-4}
Tetramers						
Primates	0.85 ± 0.02	2.85 ± 0.03	$< 2 \times 10^{-5}$	0.86 ± 0.02	4.61 ± 0.07	$< 2 \times 10^{-5}$
Vertebrates	0.89 ± 0.04	2.57 ± 0.19	$< 2 \times 10^{-5}$	0.89 ± 0.04	5.71 ± 1.17	$< 2 \times 10^{-5}$
Invertebrates	0.83 ± 0.01	1.31 ± 0.01	$< 2 \times 10^{-5}$	0.96 ± 0.02	1.53 ± 0.02	$< 2 \times 10^{-5}$
Mammals	0.68 ± 0.04	2.96 ± 0.29	$< 2 \times 10^{-5}$	0.69 ± 0.04	3.84 ± 0.46	$< 2 \times 10^{-5}$
Rodents	0.79 ± 0.02	4.57 ± 0.06	$< 2 \times 10^{-5}$	0.80 ± 0.02	11.32 ± 0.23	$< 2 \times 10^{-5}$
Plants	0.91 ± 0.01	1.85 ± 0.03	$< 2 \times 10^{-5}$	0.92 ± 0.01	2.27 ± 0.15	$< 2 \times 10^{-5}$

R_n ratios. Interestingly, the difference between the trimer clustering ratios for different taxonomic partitions in coding DNA is less pronounced than that in noncoding DNA. This indicates that coding regions are more evolutionary conserved than noncoding regions.

7. Conclusions

Observations (i)–(iii) might arise from the evolutionary pressure against clustering of repeats (except trimeric) in coding DNA. These observations are in agreement with the results on DTR, which are abundant in noncoding DNA, while they are rare in coding DNA. These differences in coding and noncoding DNA can be attributed to the fact that noncoding DNA is more tolerant to evolutionary mutational alterations than coding DNA. These findings are also consistent with the conclusions of Lio et al. [27].

For coding DNA, the observed clustering of trinucleotides could be due to specific protein structures in which amino acids cluster together (such as an alpha helix). Another possibility is that clustering of amino acids is alloted to the general problem of the stability of a native state of the folded proteins [40–43].

The strength of clustering of trimers in coding DNA relative to dimers and tetramers can be explained by the fact that insertion or deletion of a dimer or a tetramer would lead to a frame shift. Such shift in the reading frame leads in most cases to a loss of protein function, which can be lethal for the organism. On the contrary, the insertion or deletion of a trimer is equivalent to the insertion or deletion of an amino acid in the protein sequence. Such insertion or deletion, if it happens away from the functionally or structurally important sites of the protein (see [44,45]), would not affect the protein function, and hence would be tolerated by natural selection.

The source of clustering of oligomers in noncoding DNA could be the result of various duplication processes or simple repeat expansion processes [8,21], indicating that some of the neighboring oligomers evolved from the same single copy. The abundance of simple repeats in noncoding DNA contribute to the strength of long-range correlations in noncoding DNA sequences comparatively to the coding sequences (see Refs. [32,46–49] and references there in).

Acknowledgements

We would like to thank N. Broude, C. Cantor, R.S. Dokholyan, M. Frank-Kamenetskii, A. Goldberger, M. Goodman, A.Yu. Grosberg, I. Grosse, I Labat, R.N. Mantegna, C.-K. Peng, E.I. Shakhmonvich, A. Shehter, M. Simons, C. Smith, F. Starr, E.N. Trifonov, G.M. Viswanathan, G.H. Weiss, and R. Wells for fruitful discussions. This work is supported by NIH-HGP. N.V.D. is supported by NIH NRSA molecular biophysics predoctoral traineeship (GM08291-09).

References

[1] J.S. Beckmann, J.L. Weber, Survey of human and rat microsatellites, Genomics 12 (1992) 627–631.

[2] G.I. Bell, Roles of repetitive sequences, Comput. and Chem. 16 (1992) 135–143.

[3] C. Burge, A.M. Campbell, S. Karlin, Over- and under-representation of short oligonucleotides in DNA sequences, Proc. Natl Acad. Sci. USA 89 (1992) 1358–1362.

[4] B. Olaisen, M. Bekkemoen, P. Hoff-Olsen, P. Gill, Human VNTR mutation and sex, in: S.D.J. Pena, R. Chakraborty, J.T. Epplen, A.J. Jeffreys (Eds.), DNA Fingerprinting: State of the Science, Springer, Basel.

[5] R.I. Richards, G.R. Sutherland, Simple repeat DNA is not replicated simply, Nat. Genet. 6 (1994) 114–116.

[6] K. Orth, J. Hung, A. Gazdar, A. Bowcock, J.M. Mathis, J. Sambrook, Genetic instability in human ovarian cancer cell lines, Proc. Nat. Acad. Sci. USA 91 (1994) 9495–9499.

[7] A.M. Bowcock, A. Ruiz-Linares, J. Tomfohrde, E. Minch, J.R. Kidd, L.L. Cavalli-Sforza, High resolution of human evolutionary trees with polymorphic microsatellites, Nature 368 (1994) 455–457.

[8] G.R. Sutherland, R.I. Richards, Simple tandem DNA repeats and human genetic disease, Proc. Natl Acad. Sci. USA 92 (1995) 3636–3641.

[9] X. Chen, S.V. Mariappan, P. Catasti, R. Ratliff, R.K. Moyzis, A. Laayoun, S.S. Smith, E.M. Bradbury, G. Gupta, Hairpins are formed by the single DNA strands of the fragile X triplet repeats: structure and biological implications, Proc. Natl Acad. Sci. USA 92 (1995) 5199–5203.

[10] J. Jurka, C. Pethiyagoda, Simple repetitive DNA sequences from primates: compilation and analysis, J. Mol. Evol. 40 (1995) 120–126.

[11] S. Karlin, J. Mrázek, What drives codon choices in human genes, J. Mol. Biol. 262 (1996) 459–472.

[12] R.L. Stallings, Distribution of trinucleotide microsatellites in different categories of mammalian genomic sequence: implications for human genetic diseases, Genomics 21 (1994) 116–121.

[13] R.D. Wells, Molecular basis of genetic instability of triplet repeats, J. Biol. Chem. 271 (1996) 2875–2878.

[14] Y. Ionov, M.A. Peinado, S. Malkhosyan, D. Shibata, M. Perucho, Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for clonic carcinogenesis, Nature 363 (1993) 558–561.

[15] T.A. Kunkel, Slippery DNA and diseases, Nature 365 (1993) 207–208.

[16] L.A. Aaltonen, P. Peltomäki, F.S. Leach, P. Sistonen, L. Pylkkänen, J.P. Mecklin, H. Järvinen, S.M. Powell, J. Jen, S.R. Hamilton, G.M. Petersen, K.W. Kinzler, B. Vogelstein, A. de la Chapelle, Clues to the pathogenesis of familial colorectal cancer, Science 260 (1993) 812–816.

[17] S.N. Thibodeau, G. Bren, D. Schaid, Microsatellite instability in cancer of the proximal cancer, Science 260 (1993) 816–819.

[18] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, J.D. Watson, Molecular Biology of the Cell, Garland Publishing, New York, 1994.

[19] K.A. Marx, S.T. Hess, R.D. Blake, Characteristics of the large $(dA)\cdot(dT)$ homopolymer tracts in *D. discoideum* gene flanking and intron sequences, J. Biomol. Struct. Dyn. 11 1993 57– 66.

[20] G. Yagil, The frequency of two-base tracts in eukaryotic genomes, J. Mol. Evol. 37 (1993) 123–130.

[21] N.V. Dokholyan, S.V. Buldyrev, S. Havlin, H.E. Stanley, Distribution of base pair repeats in coding and noncoding DNA sequences, Phys. Rev. Lett. 79 (1997) 5182–5185.

[22] N.V. Dokholyan, S.V. Buldyrev, S. Havlin, H.E. Stanley, Model of unequal chromosomal crossing over in DNA sequences, Physica A 249 (1998) 594–599.

[23] G.I. Bell, J. Jurka, The length distribution of perfect dimer repetitive DNA is consistent with its evolution by an unbiased single step mutation process, J. Mol. Evol. 44 (1997) 414–421.

[24] D.G. Arqués, C.J. Michel, Periodicities in introns, Nucl. Acid. Res. 15 (1987) 7581–7592.

[25] A.K. Konopka, G.W. Smythers, J. Owens, J.V.Jr. Maizel, Distance analysis helps to establish characteristic motifs in intron sequences, Gene Anal. Technol. 4 (1987) 63–74.

[26] J. Mrazek, K. Kypr, Middle-range clustering of nucleotides in genomes, CABIOS 11 (1995) 195–199.

[27] P. Lio, A. Politi, S. Ruffo, M. Buiatti, Analysis of Genomic Patchiness of Haemophilus influenzae and Saccharomyces cerevisiae chromosomes, J. Theoret. Biol. 183 (1996) 455–469.

[28] A. Bunde, S. Havlin, Fractals and Disordered Systems, Springer, Berlin, 1991.

[29] D. Stauffer, A. Aharony, Introduction to Percolation Theory, Taylor & Francis, Philadelphia, 1992.

[30] P.J. Reynolds, H.E. Stanley, W. Klein, Ghost fields, pair connectedness, and scaling: exact results in one-dimensional percolation, J. Phys. A 10 (1977) L203–210.

[31] N.V. Dokholyan, S.V. Buldyrev, S. Havlin, H.E. Stanley, Distribution of dimeric tandem repeats in noncoding and coding DNA sequences, preprint, 1998.

[32] W. Li, K. Kaneko, Long-range correlations and partial $1/f^{\alpha}$ spectrum in a noncoding DNA sequence, Europhys. Lett. 17 (1992) 655–660.

[33] G.I. Bell, Evolution of simple sequence repeats, Comput. Chem. 20 (1996) 41–48.

[34] S. Kruglyak, R.T. Durrett, M.D. Schug, C.F. Aquadro, Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations, Proc. Nat. Acad. Sci. USA 95 (1998) 10774–10 778.

[35] B. Charlesworth, P. Sniegowski, W. Stephan, The evolutionary dynamics of repetative DNA in eukaryotes, Nature 371 (1994) 215–220.

[36] G. Levinson, G.A. Gutman, Slipped-strand mispairing: a major mechanism for DNA sequence evolution, Mol. Biol. Evol. 4 (3) (1987) 203–221.

[37] D. Sornette, R. Cont, Convergent multiplicative processes repelled from zero: power laws and truncated power laws, J. Phys. I France 7 (1997) 431–444.

[38] R.H.R. Stanley, N.V. Dokholyan, S.V. Buldyrev, S. Havlin, H.E. Stanley, Clustering of identical oligomers in coding and noncoding DNA sequences, Journal of Biomolecular Structure & Dynamics 17 (1999) 79–87.

[39] W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, Numerical Recepies, Cambridge University Press, Cambridge, 1989.

[40] E.I. Shakhnovich, A.M. Gutin, Implication of thermodynamics of protein folding for evolution of primary sequences, Nature 346 (1990) 773–775.

[41] V.I. Abkevich, A.M. Gutin, E.I. Shakhnovich, Impact of local and non-local interactions on thermodynamics and kinetics of protein folding, J. Mol. Biol. 252 (1995) 460–471.

[42] H. Herzel, E.N. Trifonov, O. Weiss, I. Große, Interpreting correlations in biosequences, Physica A 248 (1998) 449–459.

[43] L.A. Mirny, V. Abkevich, E.I. Shakhnovich, Universality and diversity of the protein folding scenarios: a comprehensive analysis with the aid of a lattice model, Folding Des. 1 (1996) 103–116.

[44] E.I. Shakhnovich, V.I. Abkevich, O. Ptitsyn, Conserved residues and the mechanism of protein folding, Nature 379 (1996) 96–98.

[45] N.V. Dokholyan, S.V. Buldyrev, H.E. Stanley, E.I. Shakhnovich, Molecular dynamics studies of folding of a protein-like model, Folding Des. 3 (1998) 577–587.

[46] C.-K. Peng, S.V. Buldyrev, A. Goldberger, S. Havlin, F. Sciortino, M. Simons, H.E. Stanley, Long-Range Correlations in Nucleotide Sequences, Nature 356 (1992) 168–171.

[47] W. Li, The study of correlation structure of DNA sequences: a critical review, Comput. Chem. 21 (4) (1997) 848–851.

[48] G.M. Viswanathan, S.V. Buldyrev, S. Havlin, H.E. Stanley, Quantification of DNA Patchiness using Correlation Measures, Biophys. J 72 (1997) 866–875.

[49] S.V. Buldyrev, N.V. Dokholyan, A.L. Goldberger, S. Havlin, C.-K. Peng, H.E. Stanley, G.M. Viswanathan, Analysis of DNA sequences using methods of statistical physics, Physica A 249 (1998) 430–438.