J. theor. Biol. (2000) **202,** 273–282
doi:10.1006/jtbi.1999.1052, available online at http://www.idealibrary.com on IDEAL®

Distributions of Dimeric Tandem Repeats in Non-coding and Coding DNA Sequences

NIKOLAY V. DOKHOLYAN*†‡, SERGEY V. BULDYREV*,
SHLOMO HAVLIN*§ AND H. EUGENE STANLEY*

**Center for Polymer Studies, Physics Department, Boston University, Boston, MA* 02215, *U.S.A.*,
†*Department of Chemistry and Chemical Biology, Harvard University*, 12 *Oxford Street* #42,
Cambridge, MA 02138, *U.S.A. and* §*Gonda-Goldschmied Center and Department of Physics,*
Bar-Ilan University, Ramat Gan 52900, *Israel*

(*Received on* 23 *February* 1999, *Accepted in revised form on* 8 *November* 1999)

We study the length distribution functions for the 16 possible distinct dimeric tandem repeats in DNA sequences of diverse taxonomic partitions of GenBank (known human and mouse genomes, and complete genomes of *Caenorhabditis elegans* and yeast). For *coding* DNA, we find that all 16 distribution functions are exponential. For *non-coding* DNA, the distribution functions for most of the dimeric repeats have surprisingly long tails, that fit a power-law function. We hypothesize that: (i) the exponential distributions of dimeric repeats in protein coding sequences indicate strong evolutionary pressure against tandem repeat expansion in coding DNA sequences; and (ii) long tails in the distributions of dimers in *non-coding* DNA may be a result of various mutational mechanisms. These long, non-exponential tails in the distribution of dimeric repeats in non-coding DNA are hypothesized to be due to the higher tolerance of non-coding DNA to mutations. By comparing genomes of various phylogenetic types of organisms, we find that the shapes of the distributions are not universal, but rather depend on the specific class of species and the type of a dimer.

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 interest due to their role in genome organization and evolutionary processes (Beckmann & Weber, 1992; Bell, 1992; Burge *et al*., 1992; Olaisen *et al*., 1993; Richards & Sutherland, 1994; Orth *et al*., 1994; Bowcock *et al*., 1994; Sutherland & Richards, 1995; Chen *et al*., 1995; Garcy *et al*., 1995; Jurka & Pethiyagoda, 1995; Karlin & Mrázek, 1996; Stanley *et al*., 1999). It is known that SSR constitute a large fraction of non-coding DNA and are relatively rare in protein coding sequences.

SSR are of considerable practical and theoretical interest due to their high polymorphism (Bowcock *et al*., 1994). The formation of a hairpin structure during replication (Stallings, 1994; Wells, 1996) is believed to be the cause of the *CNG* repeat expansions, which are associated with a broad variety of genetic diseases. Dimeric

‡Author to whom correspondence should be addressed. E-mail: dokh@wild.harvard.edu

0022–5193/00/040273 + 10 $35.00/0

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 (Ionov *et al.*, 1993). Similar mechanisms are attributable for other types of cancer (Orth *et al.*, 1994; Kunkel, 1993; Wooster *et al.*, 1994; Strand *et al.*, 1993; Aaltonen *et al.*, 1993; Thibodeau *et al.*, 1993). Telomeric SSR, which control DNA sequence size during replication, illustrate another crucial role of tandem repeats (Alberts *et al.*, 1994).

A study of SSR from primates, emphasizing their abundance, length polymorphism, and overall tendency to expand in different sequence contexts was reported by Jurka & Pethiyagoda (1995). The probability distribution functions for the length of special classes of repeats have been studied in many publications (see e.g. Marx *et al.*, 1993; Yagil, 1993; Stallings *et al.*, 1991; Dokholyan *et al.*, 1997, 1998). Bell & Jurka (1997) 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 (Arqués & Michel, 1987; Konopka *et al.*, 1987). 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 ("dimetric tandem repeats"). Dimeric tandem repeats are so abundant in non-coding DNA that their presence can even be observed by global statistical methods such as the power spectrum: e.g. the data of Buldyrev *et al.* (1995) indicate the presence of a peak at $(1/2)\text{bp}^{-1}$ in the power spectrum of *non-coding* DNA (corresponding to repetition of dimers) and the absence of this peak in *coding* DNA (here bp denotes base pair). This difference in the abundance of dimeric tandem repeats in coding DNA and non-coding DNA suggests that these repeats may play a special role in the organization and evolution of non-coding DNA.

The following section describes methods we use to analyse distributions of dimeric repeats. The difference between coding and non-coding DNA sequences of the distribution of dimeric tandem repeats is presented in Section 3. The relation of these distributions in coding DNA to the protein folding problem, as well as the relation to the evolutionary mechanisms are discussed in Section 4.

2. Methods

We study genomes of four different organisms: human (*Homo sapiens*), mouse (*Mus musculus*), nematode (*Caenorhabditis elegans*) 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 analyse separately all coding and non-coding 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 $(yx)_{\ell\pm 1}$ 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. Results

We find that the normalized number of repeats $N_0(\ell)$ for all six groups of dimeric tandem repeats in conding DNA and for all four analysed organisms decays rapidly with ℓ [Figs 1–4(a) and (b)]. By plotting $N_0(\ell)$ in the semi-logarithmic scale 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' correspondingly which have repeats of length larger than 10. These genes may provide either an exception from the general rule, or may

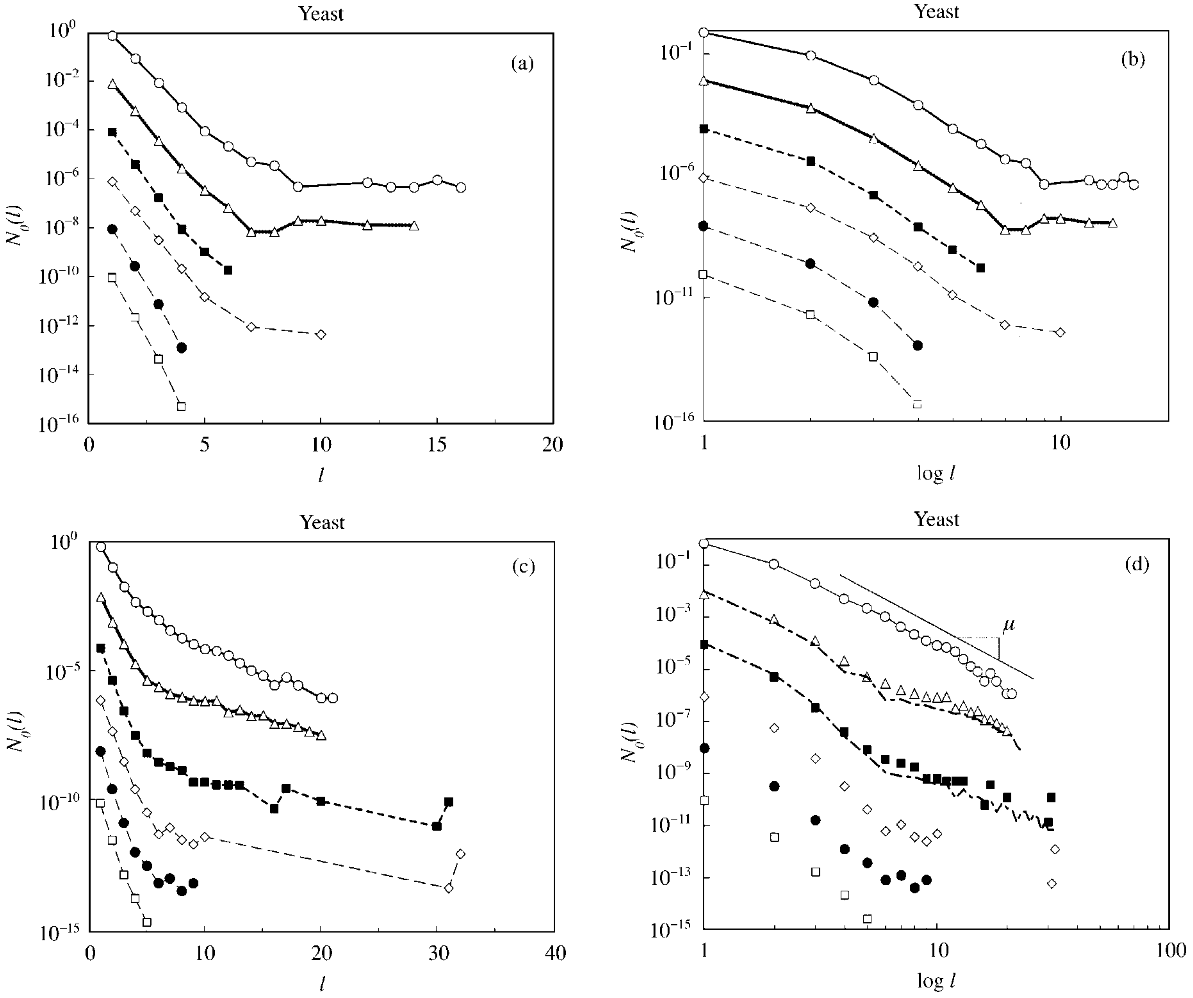

FIG. 1. The combined plot of average normalized number of repeats for six groups of dimeric tandem repeats. (○) *AA*, *TT* (*AA* or *TT*); (△) *TA*, *AT*; (■) *CA*, *AC*, *TG*, *GT*; (●) *CC*, *GG*; (◇) *GA*, *AG*, *TC*, *CT* and (□) *GC*, *CG* for the coding DNA in (a) semi-logarithmic scale and (b) double-logarithmic scale, and non-coding DNA in (c) semi-logarithmic scale and (d) double-logarithmic scale of the complete yeast genome. 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 (d) indicates a power-law function $f(x) \sim x^{-\mu}$ with $\mu = 5.3$. In (d) as an example of $P(r, \ell)$ being a function of both r and ℓ, we include the results of simulation (dot-dashed bold line) fitting the second and the third groups of repeats (see Fig. 5). The values of μ for the first five groups of repeats are 5.3, 3.2, 2.8, 2.3, 2.7 from top to bottom, fitting range is $\ell > 5$. The value of μ is undefined for *GC*, *CG* repeats in the fitting region. In all figures, for non-coding sequences the regression coefficients R of linear fit in a double-logarithmic scale range from 0.90 to 0.99 depending on the type of repeat and taxonomic partition of the GenBank. For coding DNA sequences the R coefficient of linear fit in a semi-logarithmic scale ranges from 0.91 to 1.00.

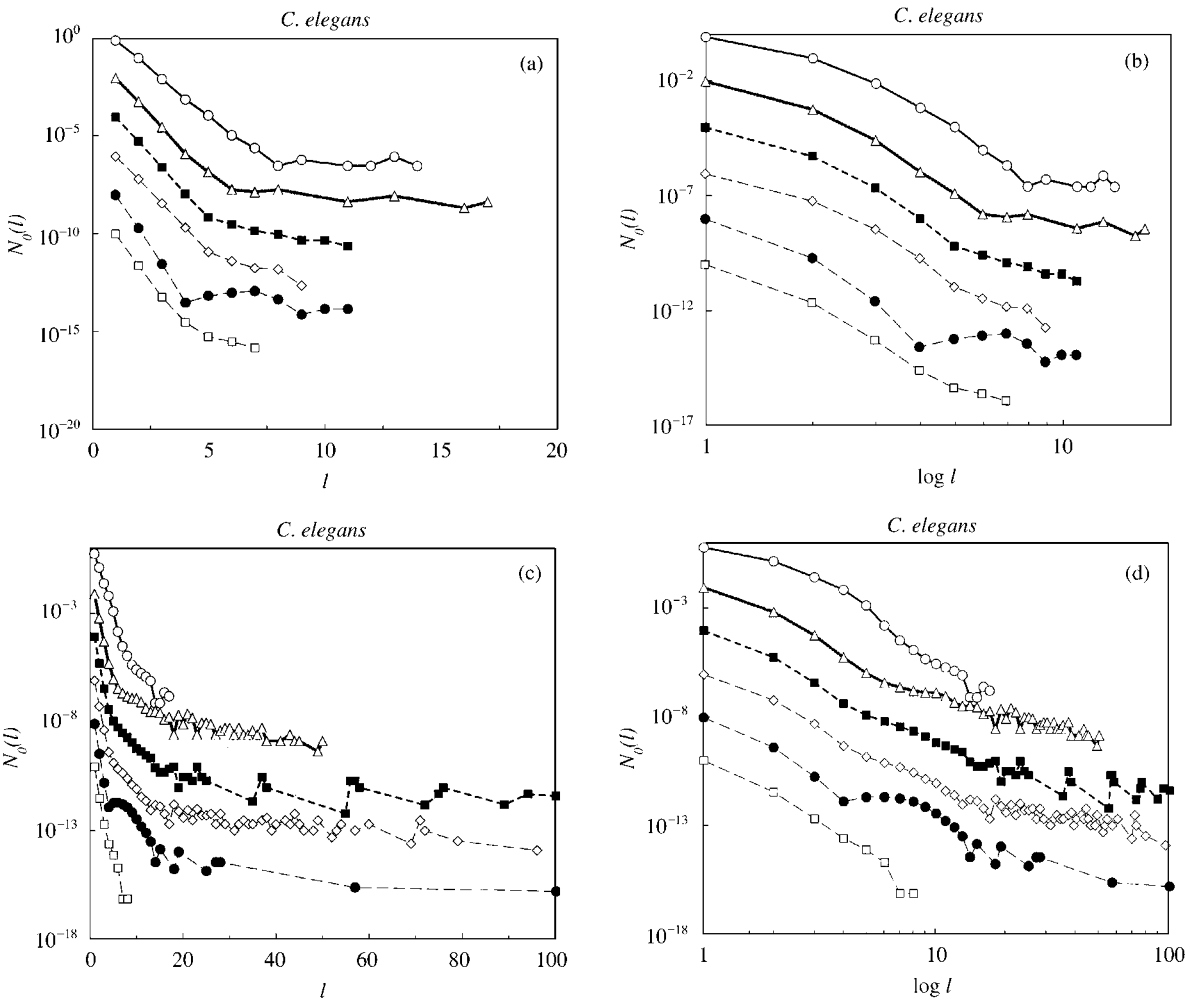

FIG. 2. The same as Fig. 1 but for complete genome of *C. elegans*. The values of μ for six groups of repeats are 7.3, 2.8, 2.6, 2.5, 3.8, 11.0 from top to bottom, fitting range is $\ell > 5$.

represent misidentified non-coding regions. It should be pointed out the CDS with long dimeric repeats, that are not verified experimentally, could be missed by gene finding algorithms, which are trained to recognize repetitive regions as non-coding. Thus, in theory, the number of long repeats in coding DNA could be underestimated in our analysis. Nevertheless, we believe that this underestimation would not significantly alter the functional form of the repeat length distribution.

On the other hand, semi-logarithmic plots of $N_0(\ell)$ for non-coding parts [see Figs 1–4(c)] are usually not straight, but display negative slope with constantly decreasing absolute value which indicates that their probability decays *less* rapidly than exponentially. This suggests that distributions $N_0(\ell)$ can be better described by a power law

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

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 1–4(d). 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 non-coding DNA (introns and intergenic sequences).

There are two exceptions:

(i) We find that in some cases, the double logarithmic plots for non-coding regions have long

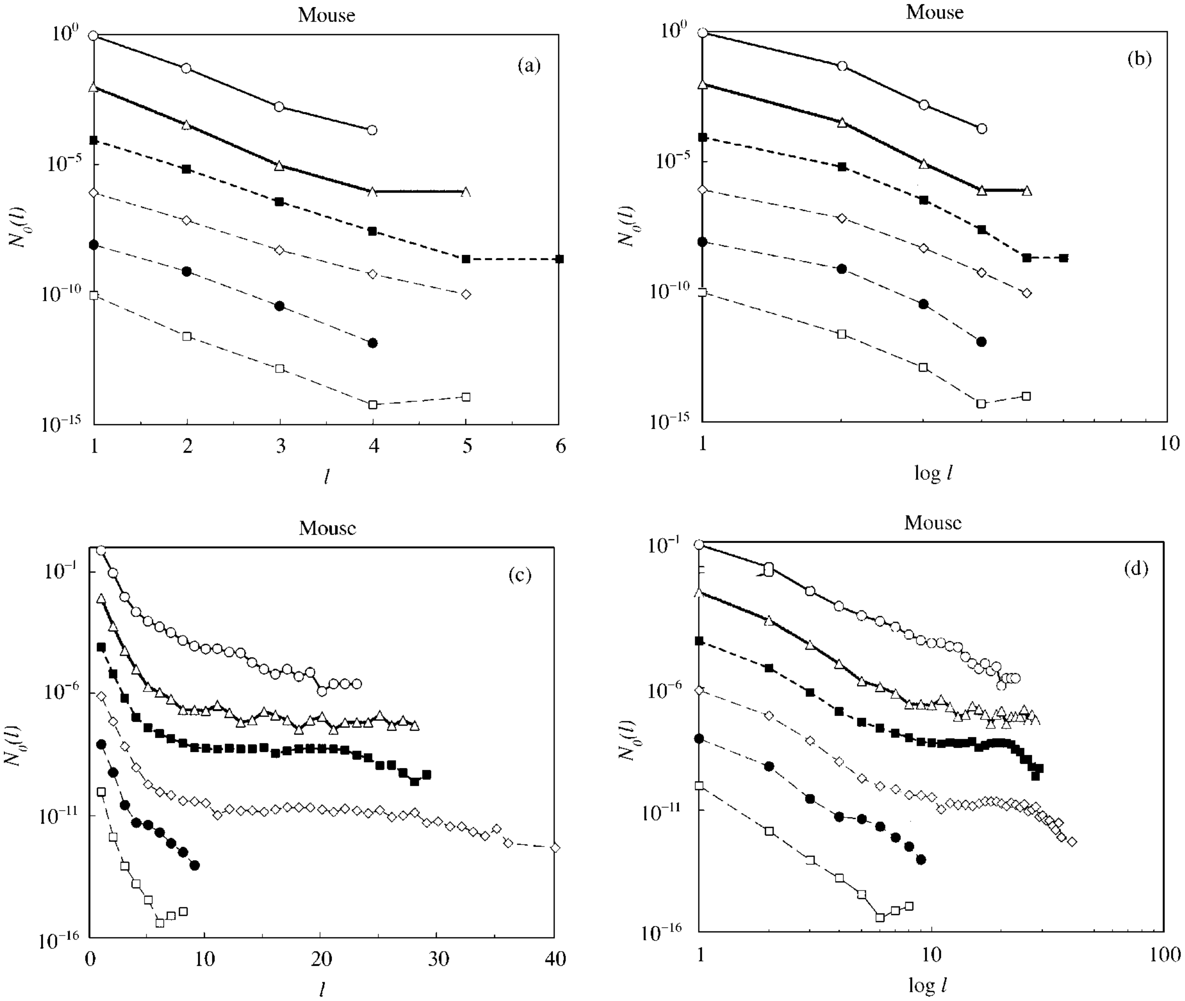

FIG. 3. The same as Fig. 1 but for known mouse genome. The values of μ for first five groups of repeats are 4.2, 1.8, 1.9, 1.9, 6.4 from top to bottom, fitting range is $\ell > 5$. The value of μ is undefined for *GC*, *CG* repeats in the fitting region.

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 non-coding sequences of the dimers *CA–AC*, *TG–GT*, and *GA–AG*, *TC–CT* in mouse and human DNA [see Figs 3(d) and 4(d)]. This effect is not observed in *C. elegans* and yeast.

(ii) We also observe that all organisms lack *CG–GC* repeats in their intron and non-coding DNA [see Figs 1–4(b) and (c)].

It is interesting to note that the maximal length of dimeric tandem repeats in human and mouse DNA does not exceed 40 copies. In contrast, in *C. elegans*, repeats can be as large as 100 copies. While this fact may indicate the presence of specific mechanisms that prevents dimeric repeat expansion in mammals, it may also be due to the fact that we analyse the entire genome of *C. elegans*, while human and mouse data are restricted to the gene-rich regions.

4. Discussion

4.1. SIGNIFICANCE OF THE OBSERVED DISTRIBUTIONS

The difference in distributions of dimeric tandem repeats in coding and non-coding DNA is drastic [cmp. Figs 1–4(a, b) and (c, d)]. In non-coding DNA we find long chains, up to 100 repetitions, of tandem repeats like "*TGTG*..." (in the intergenic sequences of *C. elegans* (Fig. 2). Note that if one assumes that a DNA sequence is

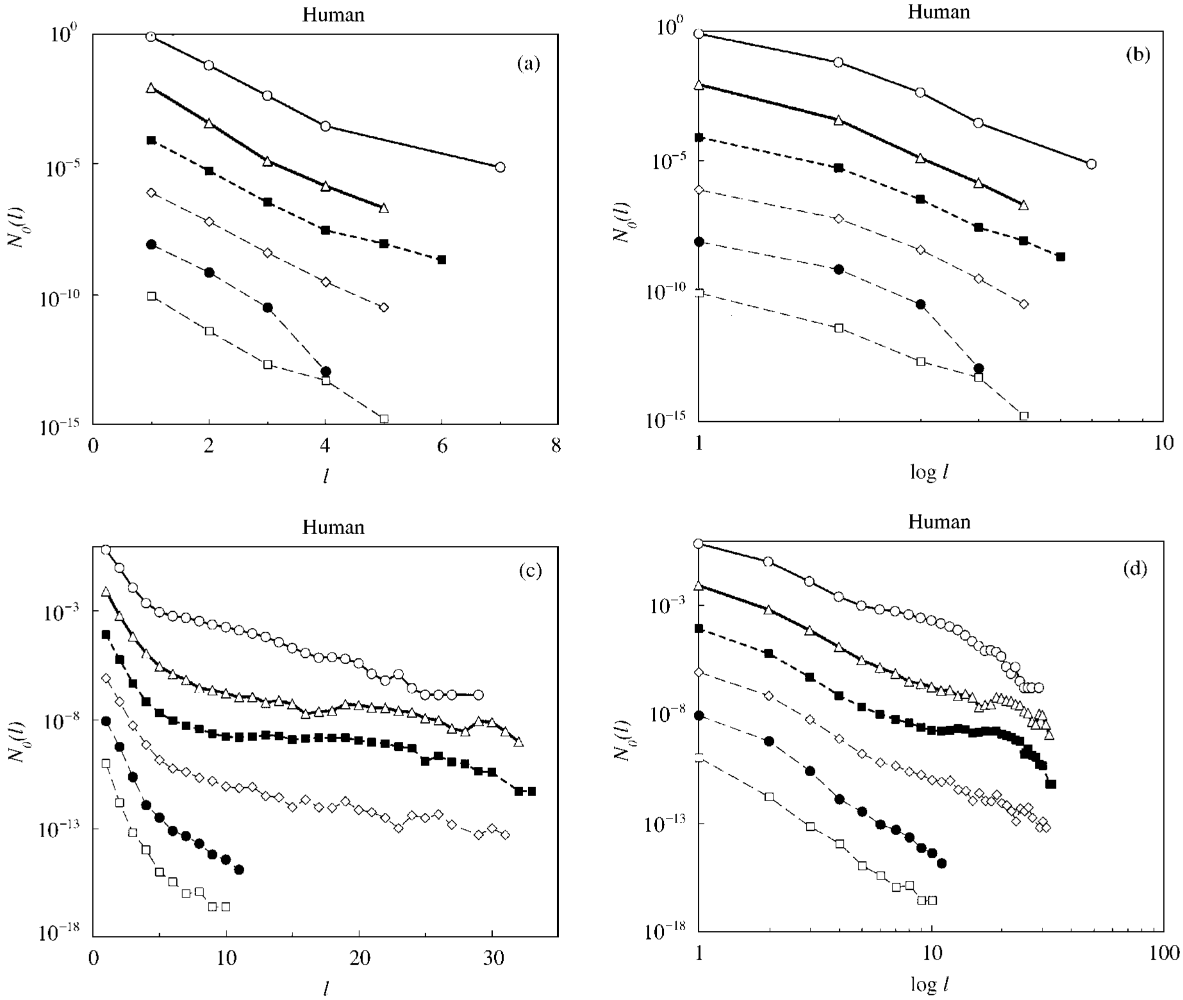

FIG. 4. The same as Fig. 1 but for known human genome. The values of μ for these six groups of repeats are 5.6, 3.3, 3.2, 4.1, 6.7, 5.4 from top to bottom, fitting range is $\ell > 5$.

uncorrelated or short-range correlated, then the estimated probability of such a repeat is of the order of 10^{-118}, which makes it highly improbable to find it in any organism during the entire evolution. However, this argument fails if we assume that distribution of repeats fits a power law with $\mu = 4$. Then, the probability to find a 100 repetitions of TG is of the order of 10^{-8}, which is of the order of the inverse size of the *C. elegans* genome bank. This is the reason why any model resulting in exponential length distributions will be unable to explain long repeats (see critical review by Li, 1997).

4.2. RELATION TO TRIMERS

The absence of lengthy dimeric tandem repeats in coding DNA is related to the 3 bp length of a codon. Since 3 is not a multiple of 2, the repetition in amino acid sequence does not correspond to the repetition of dimers. Only very specific pairs of amino acids, such as *HisThr* (*CACACA*), *ArgGlu* (*AGAGAG*), *ArgAla* (*CGCGCG*), correspond to dimeric tandem repeats. In this case, dimeric repeats of amino acids correspond to those of DNA. Even simple repeats, such as $(AA)_\ell$ and $(TT)_\ell$, are much more abundant in non-coding DNA. This observation probably indicates that there is strong evolutionary pressure against long stretches of amino acid repeats in proteins. It was proposed (Shakhnovich & Gutin, 1990; Le *et al.*, 1997) that, in order to fold into a native state, proteins should have statistical properties close to those of a random sequence. Our findings are in complete agreement with the absence of long amino acid repeats in globular proteins.

4.3. CAN MUTATIONS EXPLAIN NON-EXPONENTIAL DISTRIBUTIONS OF REPEATS?

Recently, several mechanisms of simple sequence repeat expansion have been proposed (Bell, 1992, 1996; Li & Kaneko, 1992; Richards & Sutherland, 1994; Orth *et al*., 1994; Chen *et al*., 1995; Wells, 1996; Bell and Jurka, 1997; Dokholyan *et al*., 1997, 1998; Kruglyak *et al*., 1998). The model, proposed recently to explain power-law distributed repeats (Dokholyan *et al*., 1997) can produce power-law distributed repeats with any given exponent μ (for details see Dokholyan *et al*., 1998).

The mechanism proposed in Dokholyan *et al*. (1997, 1998) is based on random multiplicative processes, which can reproduce the observed non-exponential distribution of repeats. The increase of decrease of repeat length can occur due to unequal crossover (Alberts *et al*., 1994; Charlesworth *et al*., 1994) or slippage during replication (Richards & Sutherland, 1994; Wells, 1996; Levinson & Gutman, 1987; Schlötterer & Tautz, 1992).

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 Sornette & Cont (1997) and in Dokholyan *et al*. (1997, 1998), 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 (Levinson & Gutman, 1987; Wells, 1996; Dokholyan *et al*., 1997, 1998), 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 $\ell \geqslant 5$.

It was shown by computer simulations (Dokholyan *et al*., 1997) that imposing the length constraints on the mutational rates $P(r, \ell)$ produces better fit of experimental data [see Fig. 1(d)]. For example, the probability distribution function $P(r, \ell)$, has been chosen as shown in Fig. 5 to fit distributions of TA, AT and CA, AC, TG, GT repeats in yeast [Fig. 1(d)]. 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 Kruglyak *et al*. (1998), 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 $r = \ell'/\ell, \ell$ is the original length and ℓ' is a repeat length after a time interval during which several stepwise mutations can occur.

A convenient property of the model proposed in Dokholyan *et al*. (1997) 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 Fig. 1 and also Dokholyan *et al*., 1998). The "plateau" in the distributions of dimeric repeats, discussed in the Section 3, can also be explained by the mutational rate variability and can be fit by the model.

It is possible that the existence of long TG repeats and a lack of CG repeats in vertebrates is related to methylation (Alberts *et al*., 1994). In germ cells, the cytosine nucleotide C in the sequence CG can be covalently modified to 5-*methylcytosine* (5-*methyl* C). The methylated C nucleotide undergoes accidental de-amination which cannot be repaired back to the previous state of C, but gets modified into a T nucleotide. Therefore, there is a strong tendency for methylated CG to mutate to TG. Three out of four CGs have been lost during evolution due to this mechanism (Alberts *et al*., 1994).

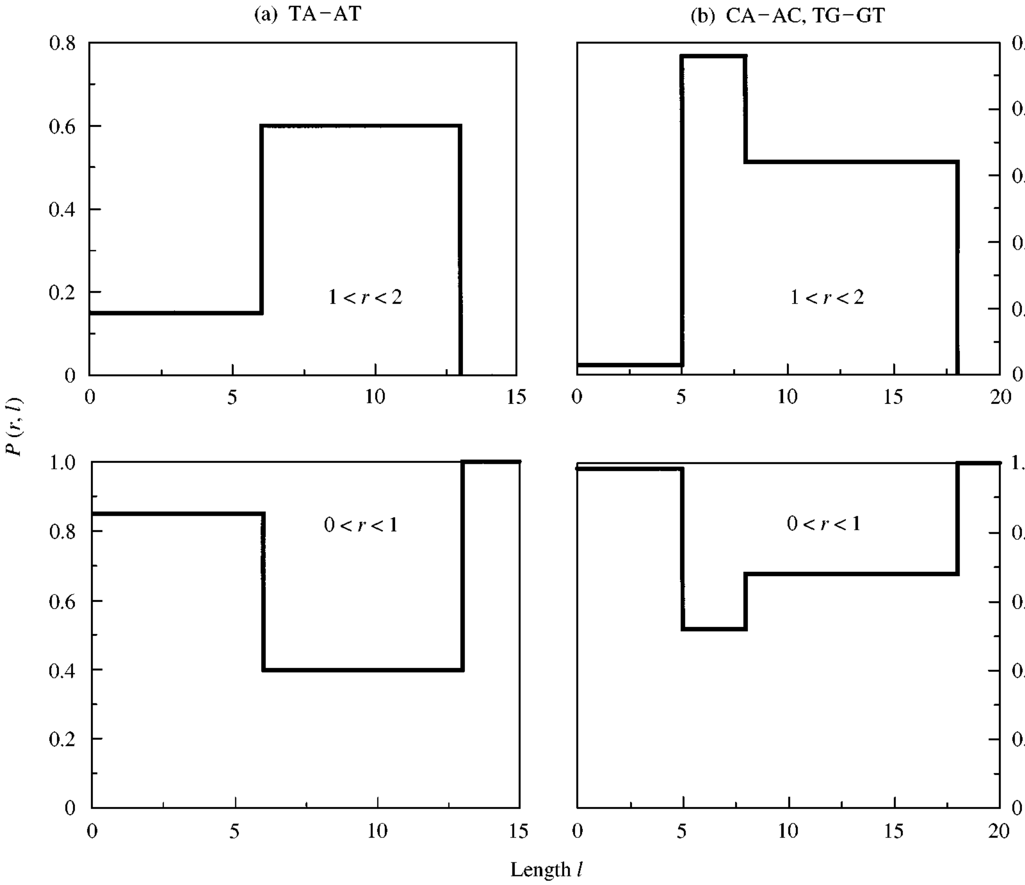

FIG. 5. As an example of $P(r, \ell)$ being a function of both r and ℓ, we use following $P(r, \ell)$ to fit *TA*, *AT CA*, *AC*, *TG*, *GT* repeats in non-coding yeast DNA (see Fig. 1). (a) for *TA*, *AT* repeats, $P(r, \ell)$ depends on ℓ as a step function: for $1 < r \leqslant 2$: $P(r, \ell) = 0.15$, 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.

Another reason for a lack of CG repeats may be due to their ability to form Z-DNA in the negatively supercoiled regions during transcription (Rahmouni & Wells, 1989, 1992; Kladde *et al.*, 1994; Krasilnikov *et al.*, 1999).

4.4. POSSIBLE ROLE OF REPEATS

It is reasonable to assume that the power-law distributed repeats are the sign of the random multiplicative processes that are intrinsic to DNA structure. The existence of the plateaus in the distributions of *TG* repeats indicate deviation power-law distribution and may be attributed to a specific function of these repeats, such as to induce or repress transcription and to stimulate recombination (Kladde *et al.*, 1994).

Since long dimeric repeats are known to contribute to genomic instability, there are mechanisms that prevent them from uncontrollable expansion, such as *MSH2* mismatch-repair enzyme. Hence, we do not expect mutations to accumulate at a maximal rate unless there are misfunctions of repair mechanisms, which have been proven to be a cause of several types of cancers (Ionov *et al.*, 1993; Orth *et al.*, 1994; Kunkel 1993; Wooster *et al.*, 1994; Strand *et al.*, 1993; Aaltonen *et al.*, 1993; Thibodeau *et al.*, 1993).

5. Conclusion

We find that the statistical properties of dimeric tandem repeats differ in coding and non-coding parts of DNA. For exons, we find that length distributions of dimeric tandem repeats are exponential and, thus, are consistent with uncorrelated or short-range correlated amino acid sequences. On the other hand, the length distributions of dimeric repeats in introns and intergenic regions usually deviate strongly from an exponential function and can, in some cases, be fit by a power-law function. We also find that the distribution of AC–CA, TG–GT repeats in vertebrates have plateaus in the range $10 < \ell < 30$ and that the distributions of CC, CG, GC, and GG repeats decay much faster than other repeats which include weakly bonded bp.

Non-exponential distributions of SSR can be explained if one assumes a random multiplicative process for the mutation of the repeat length, i.e. each mutation leads to a change of repeat length by a random factor with a certain distribution. Such a process may take place due to errors in replication (Wells, 1996) or unequal chromosomal cross-over.

SSR expansion in the coding regions leads to a loss of protein functionality [as e.g. in Huntington's disease (Wells, 1996)] and to the extinction of the genotype. Thus, SSR distribution in protein coding sequences remain exponential due to Darwinian evolutionary pressure. Moreover, conservation of protein coding sequences (Pande *et al*., 1990) and, thus, lack of expansion dynamics of SSR in the coding regions of DNA, are probably related to the problem of protein folding. Monte-Carlo simulations of protein folding on the cubic lattice suggest that the statistical properties of the sequences that fold into a native state resemble those of random sequences (Shakhnovich & Gutin, 1990; Li *et al*., 1997), i.e. they do not have long amino acid repeats.

The higher tolerance of the non-coding DNA to various mutations, especially to mutations involving the change of DNA length—e.g. duplication, insertion of transposable elements, and SSR expansion—lead to the statistical features, that are distinct from the coding DNA Viswanathan *et al*., 1997).

We would like to thank R. S. Dokholyan, Dr M. Frank-Kamenetskii, Dr C.-K. Peng, R. Stanley, Dr G. H. Weiss, and Dr R. Wells for fruitful discussions. This work is supported by NIH-HGP, N.V.D. acknowledges support by NIH NRSA molecular biophysics predoctoral traineeship GM08291-09 and by NIH postdoctoral fellowship 1F32 GM20251-01.

REFERENCES

AALTONEN, L. A., PELTOMÄKI, P., LEACH, F. S., SISTONEN, P., PYLKKÄNEN, L., MECKLIN, J. P., JÄRVINEN, H., POWELL, S. M., JEN, J., HAMILTON, S. R., PETERSON, G. M., KINZLER, K. W., VOGELSTEIN, B. & DE LA CHAPELLE, A. (1993). Clues to the pathogenesis of familial colorectal cancer. *Science* **260,** 812–816.

ALBERTS, B., BRAY, D., LEWIS, J., RAFF, M., ROBERTS, K. & WATSON, J. D. (1994). *Molecular Biology of the Cell*. New York: Garland Publishing.

ARQUÉS, D. G. & MICHEL, C. J. (1987). Periodicities in introns. *Nucl. Acid. Res.* **15,** 7581–7592.

BECKMANN, J. S. & WEBER, J. L. (1992). Survey of human and rat microsatellites. *Genomics* **12,** 627–631.

BELL, G. I. (1992). Roles of repetitive sequences. *Comput. Chem.* **16,** 135–143.

BELL, G. I. (1996). Evolution of simple sequence repeats. *Comput. Chem.* **20,** 41–48.

BELL, G. I. & JURKA, J. (1997). The length distribution of perfect dimer repetitive DNA is consistent with its evolution by an unbiased single step mutation process. *J. Mol. Evol.* **44,** 414–421.

BOWCOCK, A. M., RUIZ-LINARES, A., TOMFOHRDE, J., MINCH, E., KIDD, J. R. & CAVALLI-SFORZA, L. L. (1994). High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* **368,** 455–457.

BULDYREV, S. V., GOLDBERGER, A. L., HAVLIN, S., MANTEGNA, R. N., MATSA, M. E., PENG, C.-K., SIMONS, M. & STANLEY, H. E. (1995). Long-range correlation properties of coding and noncoding DNA sequences: GenBank analysis. *Phys. Rev. E* **51,** 5084–5091.

BURGE, C., CAMPBELL, A. M. & KARLIN, S. (1992). Over and under-representation of short oligonucleotides in DNA sequences. *Proc. Nat. Acad. Sci. U.S.A.* **89,** 1358–1362.

CHARLESWORTH, B., SNIEGOWSKI, P. & STEPHAN, W. (1994). The evolutionary dynamics of repetative DNA in eukaryotes. *Nature* **371,** 215–220.

CHEN, X., MARIAPPAN, S. V., CATASTI, P., RATLIFF, R., MOYZIS, R., K. LAAYOUN, A., SMITH, S. S., BARDBURY, E. M. & GUPTA, G. (1995). Hairpins are formed by the single DNA strands of the fragile X triplet repeats: structure and biological implications. *Proc. Nat. Acad. Sci. U.S.A.* **92,** 5199–5203.

DOKHOLYAN, N. V., BULDYREV, S. V., HAVLIN, S. & STANLEY, H. E. (1997). Distribution of base pair repeats in coding and non-coding DNA sequences. *Phys. Rev. Lett.* **79,** 5182–5185.

DOKHOLYAN, N. V., BULDYREV, S. V., HAVLIN, S. & STANLEY, H. E. (1998). Model of unequal chromosomal crossing over in DNA sequences. *Physica* **A 249,** 594–599.

HUNTINGTON'S DISEASE COLLABORATIVE RESEARCH GROUP (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72,** 971–983.

IONOV, Y., PEINADO, M. A., MALKHOSYAN, S., SHIBATA, D. & PERUCHO, M. (1993). Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for clonic carcinogenesis. *Nature* **363,** 558–561.

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

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

Kladde, M. P., Kohwi, Y., Kohwi-Shigematsu, T. & Gorski, J. (1994). The non B-DNA structure of $d(CA/TG)_n$ dimers from that of Z-DNA. *Proc. Nat. Acad. Sci. U.S.A.* **91,** 1898–1902.

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

Krasilnikov, A. S., Podtelezhnikov, A., Vologodskii, A. & Mirkin, S. M. (1999). Large-scale effects of transcriptional DNA supercoiling *in vivo*. *J. Mol. Biol.* **292,** 1149–1160.

Kruglyak, S., Durrett, R. T., Schug, M. D. & Aquadro, C. F. (1998). Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc. Nat. Acad. Sci. U.S.A.* **95,** 10774–10778.

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

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

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

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

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

Olaisen, B., Bekkemoen, M., Hoff-Olson, P. & Gill, P. (1993). Human VNTR mutation and sex. In: *DNA Fingerprinting*: *State of the Science* (Pena, S. D. J., Chakraborty, R., Epplen, J. T. & Jeffreys, A. J., eds). Basel: Springer-Verlag.

Orth, K., Hung, J., Gazdar, A., Bowcock, A., Mathis, J. M. & Sambrook, J. (1994). Genetic instability in human ovarian cancer cell lines. *Proc. Nat. Acad. Sci. U.S.A.* **91,** 9495–9499.

Pande, V. S., Grosberg, A. Yu. & Tanaka, T. (1990). Non-randomness in protein sequences: evidence for a physically driven stage of evolution? *Proc. Nat. Acad. Sci. U.S.A.* **91,** 12 972.

Rahmouni, A. R. & Wells, R. D. (1989). Stabilization of Z-DNA *in vivo* by localized supercoiling. *Science* **246,** 358–363.

Rahmouni, A. R. & Wells, R. D. (1992). Direct evidence for the effect of the transcription of local DNA supercoiling *in vivo*. *J. Mol. Biol.* **223,** 131–144.

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

Shakhnovich, E. I. & Gutin, A. M. (1990). Implications of thermodynamics of protein folding for evolution of primary sequences. *Nature* **346,** 773–775.

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

Stallings, R. L., Ford, A. F., Nelson, D., Torney, D. C., Hildebrand, C. E. & Moyzis, R. K. (1991). Evolution and distribution of $(GT)_n$ repetitive sequences in mammalian genomes. *Genomics* **10,** 807–815.

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

Stanley, R. H. R., Dokholyan, N. V., Buldyrev, S. V., Havlin, S. & Stanley, H. E. (1999). Clumping of identical oligonucleotides in coding and non-coding DNA sequences. *J. Biomol. Struct. Dyn.* **17,** 79–87.

Strand, M., Prolla, T. A., Liskay, R. M. & Petes, T. D. (1993). Destabilization of tracts of simple repeatative DNA in yeast by mutations affecting DNA. *Nature* **365,** 274–276.

Sutherland, G. R. & Richards, R. I. (1995). Simple tandem DNA repeats and human genetic disease. *Proc. Nat. Acad. Sci. U.S.A.* **92,** 3636–3641.

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

Viswanathan, G. M., Buldyrev, S. V., Havlin, S. &. Stanley, H. E. (1997). Quantification of DNA patchiness using long-range correlation measures. *Biophys. J.* **72,** 866–875.

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

Wooster, R., Cleton-Jansen, A.-M., Collins, N., Mangion, J., Cornelis, R. S., Cooper, C. S., Gusterson, B. A., Ponder, B. A. J., von Deimling, A., Wiestler, O. D., Cornelisse, C. J., Devilee, P. & Stratton, M. R. (1994). Instability of short tandem repeats (microsatellites) in human cancers. *Nat. Genet.* **6,** 152–156.

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