Clustering of Identical Oligomers in Coding and Noncoding DNA Sequences

Rachel H. R. Stanley[1,2], Nikolay V. Dokholyan[1], Sergey V. Buldyrev[1],
Shlomo Havlin,[1,3] and H. Eugene Stanley[1]
[1] *Center for Polymer Studies and Physics Dept., Boston University, Boston, MA 02215 USA*
[2] *Chemistry Dept., MIT, Cambridge, MA 02139 USA*
[3] *Gonda-Goldschmied Center and Physics Dept., Bar-Ilan University, Ramat Gan, ISRAEL*

We develop a quantitative method for analyzing repetitions of identical short oligomers in coding and noncoding DNA sequences. We analyze sequences presently available in the GenBank separately for primate, mammal, vertebrate, rodent, invertebrate and plant taxonomic partitions. We find that some oligomers "cluster" more than they would if randomly distributed, while other oligomers "repel" each other. To quantify this degree of clustering, we define clustering measures. We find that *(i)* clustering significantly differs in coding and noncoding DNA; *(ii)* in most cases, monomers, dimers and tetramers cluster in noncoding DNA but appear to repel each other in coding DNA. *(iii)* The degree of clustering for different sources (primates, invertebrates, and plants) is more conserved among these sources in the case of coding DNA than in the case of noncoding DNA. *(iv)* In contrast to other oligomers, we find that trimers always prefer to cluster. *(v)* Clustering of each particular oligomer is conserved within the same organism.

I. INTRODUCTION

Recently there have been reports linking certain neurological diseases, such as Huntington's disease, fragile X-linked mental retardation, and myotonic dystrophy, with trinucleotide expansions — long repetitions of identical trinucleotides in the coding regions of certain genes [1–3]. For a review on the role of trinucleotide repeats in neurological diseases, see Ref. [4]. Other studies have noticed long tandem repeats of identical dinucleotides in noncoding regions of the genome [5–7]. Identical mono-, di-, tri- or tetranucleotides tandemly repeated, also known as microsatellites, have been extensively analyzed. Since microsatellites were first shown to aid genetic mapping [8], they have become primary genetic markers [9]. More recently, studies have used microsatellites to compare evolution among different species [10]. At some loci microsatellites are so polymorphic that they can be used for DNA fingerprinting [11].

Quantitative studies of microsatellites include analysis of runs of single nucleotides [12,13]. Dinucleotides have been studied in terms of nearest neighbors [14,15] and relative frequencies [6,7,16]. Trinucleotides have been examined in terms of biased distributions [17]. Tandemly repeated pentamers have been studied in [18]. Also, a comprehensive study of the average length of simple repeats of units of 1–6 nucleotides was compiled [19]. An analysis of clustering of nucleotides has been done by Mrazek and Kypr [20] and by Lio et al. for *Haemophilus influenzae* and *Saccharomyces cerevisiae* chromosomes [21].

Interest in the nucleotide patterns in DNA (such as simple sequence repeats) is growing due to its direct correspondence to evolutionary processes. The difference in nucleotide patterns in coding and noncoding DNA reflects difference in the evolutionary pressure in various functional parts of DNA. Recent studies of distributions of dimeric tandem repeats (DTR) in DNA sequences reveal a significant difference between coding and noncoding regions [6,7]. It was found that some of the DTR in noncoding DNA have power-law distribution functions. On the contrary, all the DTR distribution functions in coding DNA are exponential, which implies that they are either randomly distributed or short-range correlated. DTR are one of many examples of complex patterns in DNA.

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. Inspired by percolation theory [22–24], we calculate the mean length (defined below) of repetitions of oligomers, and the expected length of repetitions if the oligomers were randomly distributed, which we use 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 scrambled) or "repel" (repeat less than they would if their order were randomly scrambled). In such a way we can understand if oligomers in DNA tend to aggregate or segregate.

We systematically compare clustering in coding and noncoding DNA for different organisms: primate, mammal, vertebrate, rodent, invertebrate and plant taxonomic partitions of GenBank release 104.0. It is possible that differences in the patterns of repetitions in coding and noncoding DNA *(i)* can furnish ways to classify unknown sequences as coding or noncoding and *(ii)* can shed light on the dynamics of evolution of various regions of DNA.

II. METHOD: RATIO ANALYSIS

We quantify the repetitions of oligomers by applying the ratio analysis method separately to all possible reading frames (RF). Coding DNA has three RF, one of which is the correct one. Noncoding DNA, however, does not have a RF, so there is no *a priori* reason to select a given one. Here, we report only our results using the biological RF for coding DNA and a single RF randomly chosen for noncoding DNA.

First, we count the actual number of occurrences of the repetitions for each oligomer. We then 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 based on the frequency of oligomers in the GenBank sequence. Thus, to preserve the frequencies of oligomers, we scramble the order of the oligomers in the control.

For the control, we can compute the probability P_ℓ that a given oligomer belongs to a "cluster" (aggregate) [22,23] of *exactly* ℓ repetitions in an uncorrelated random sequence[1]

$$P_\ell = \ell p_i^\ell (1 - p_i)^2 \tag{1}$$

where p_i is the frequency of a particular oligomer. By multiplying each of these probabilities (a distinct probability for each of the oligomers) by L, the total number of oligomers in the GenBank sequence, we find the number of oligomers that belong to clusters of size ℓ. Thus, for random uncorrelated sequences, the expected mean number of clusters $N_i(\ell)$ of size ℓ (the control) is given by

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

We compute the number of repeats of length ℓ of a given repeat in different partitions of the GenBank: $N_i(\ell)$, where $i = 1, \ldots, M$ 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

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

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[2]). Noncoding sequences we identify as those that are not explicitly specified as exons in the GenBank database. 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 entries that exceed in length 10^4 bp. This reduces the redundancy of the GenBank. The total length and the number of sequences analyzed in coding and noncoding regions is reported in Table I.

Here 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.

[1] For the Markov sequence consult Appendix C.

[2] We do not find any significant difference between clustering properties of the untranslated $5'$ and $3'$ ends and noncoding DNA sequences. In addition, we perform the clustering ratio analysis of the expressed sequence tags (EST) database, which are mainly taken from the untranslated $5'$ and $3'$ ends, and find that their clustering properties are similar to those of noncoding sequences.

(ii) We also define the "weight" average (see e. g. [24]):

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

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.

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 (see Appendix A for the derivation). For such an uncorrelated random sequence we find

$$\langle \ell \rangle_n^{\rm th} = \frac{1}{1 - \sum_{i=1}^{M} p_i^2}, \tag{7}$$

where p_i is the frequency of each oligomer, and

$$\langle \ell \rangle_w^{\rm th} = 1 + 2 \sum_{i=1}^{M} \frac{p_i^2}{1 - p_i}. \tag{8}$$

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

$$R_n \equiv \frac{\langle \ell \rangle_n - 1}{\langle \ell \rangle_n^{\rm th} - 1} \qquad \text{and} \qquad R_w \equiv \frac{\langle \ell \rangle_w - 1}{\langle \ell \rangle_w^{\rm th} - 1}. \tag{9}$$

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, computed in Appendix B. In Table II, we present the relative clustering ratios R_n and R_w.

III. RESULTS AND DISCUSSION

We compare the ratios of the observed values of the average length $\langle l \rangle_n$ of oligomers[3] (monomers, dimers, trimers, and tetramers) and their weight average $\langle l \rangle_w$ to the theoretically predicted for a randomly scrambled sequence. We consider primate, vertebrate, invertebrate, mammal, rodent, and plant taxonomic partitions of GenBank release 104[4]. The complete results for clustering ratio values and for the error bars of these values are presented in Table II. 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 II. 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 [25]). If p is close to 1, then the two distributions are drawn from the same distribution with the probability 1. If p is close to 0, then these distributions are taken from two different distributions with the probability $(1-p) \to 1$. In Table II we also present the p-values. The errors which are due to the finite length of the sequences are negligible (see Appendix B). We find:

(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 $2 \cdot 10^{-5}$.

(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.

[3] We are unable to obtain statistically significant results for the oligomers, longer than tetramers. Hence, we omit the results for pentamers, hexamers, etc. in the present report.

[4] We also considered the complete genome of *Escherichia coli*, however we found clustering in neither coding nor noncoding DNA.

(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.

(iv) The clustering ratios for the trimers for all organisms show strong clustering of the trimers in both coding and noncoding DNA.

(v) 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.

Interestingly, the difference between the clustering of trimers in coding DNA is less pronounced than in noncoding DNA. For primates and mammals the Kolmogorov-Smirnov p-values for the R_n ratio are of the order of 1 (Table II), which indicates that one cannot distinguish between coding and noncoding DNA based only on R_n ratios.

Observations *(i)* – *(v)* might arise from the evolutionary pressure against clustering of repeats (except trimeric) in coding DNA. These observations are in agreement with recent work [7,15,26], where the dimeric tandem repeats (DTR) were studied and it was found that DTR are abundant in noncoding DNA, while they are rare in coding DNA. A proposed mechanism of expansion [6,27], based on the different mutational mechanisms, and the difference in length distributions of DTR in coding and noncoding DNA was attributed to the fact that noncoding DNA is more tolerant to evolutionary mutational alterations than coding DNA. These findings are also consistent with the work of Lio et al. [21]. An in-depth discussion of how the mutation processes affect the distribution of dimers can be found in [7].

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 [28–31].

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 RF 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 [32,33]), 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 [5,6], indicating that some of the neighboring oligomers evolved from the same single copy.

We also calculate the clustering measures for each individual oligomer, which we define the same way as in Eqs. (4) – (9), except that the summation over all types of oligomers ($i = 1, 2, ..., M$) is omitted in calculations of number and weight average values, and clustering ratios. Hence,

$$R_{n,i} \equiv \frac{\langle \ell \rangle_{n,i} - 1}{\langle \ell \rangle_{n,i}^{\rm th} - 1} = \frac{\sum_{\ell=1}^{\infty} \ell_i N_i(\ell_i) / \sum_{\ell=1}^{\infty} N_i(\ell_i) - 1}{1/(1-p_i) - 1} \,, \tag{10}$$

and

$$R_{w,i} \equiv \frac{\langle \ell \rangle_{w,i} - 1}{\langle \ell \rangle_{w,i}^{\rm th} - 1} = \frac{\sum_{\ell=1}^{\infty} \ell_i^2 N_i(\ell_i) / \sum_{\ell=1}^{\infty} \ell_i N_i(\ell_i) - 1}{(1+p_i)/(1-p_i) - 1} \,. \tag{11}$$

The theoretical values for $\langle \ell \rangle_{n,i}^{\rm th} = 1/(1-p_i)$ and $\langle \ell \rangle_{w,i}^{\rm th} = (1+p_i)/(1-p_i)$ (see [24]) are computed similarly to Eq. (7) and (8).

(vi) We find that the clustering ratios for each individual oligomer is conserved for each organism, i. e. the standard deviation of the clustering measures is around a few per cent. To illustrate this observation we report the clustering ratio values for dimers in primates in Table III[5]. This observation indicates that the clustering ratio can characterize the unique property of the DNA sequences to cluster and can be utilized in further studies of aggregates of oligomers in DNA sequences. For example, different clustering ratios of various dimers can suggest different mutation rates, specific for each dimer and organism.

[5] The data for the clustering ratio values for other oligomers and taxonomic partitions of the GenBank are consistent with this statement.

Acknowledgments

We wish to thank N. Goodman, A. Shehter, F. W. Starr and especially C. Smith for helpful discussions. We also thank referees for a number of helpful suggestions. We acknowledge NIH for financial support. N. V. D. is supported by NIH NRSA molecular biophysics predoctoral traineeship (GM08291-09).

APPENDIX A: DERIVATION OF EQS. (7) AND (8)

To derive Eq. (7) let us start from the definition Eq. (4) of the length average $\langle \ell \rangle_n$:

$$\langle \ell \rangle_n = \frac{L}{\sum_{\ell=1}^{\infty}\sum_{i=1}^{M} N_i^0(\ell)} = \frac{1}{\sum_{\ell=1}^{\infty}\sum_{i=1}^{M} N_i^0(\ell)/L}\,, \tag{A1}$$

where $N_i^0(\ell)$, the number of oligomers of type i in a sequence of length L, is determined by the Eq. (2). Thus,

$$\langle \ell \rangle_n = \frac{1}{\sum_{i=1}^{M}\sum_{\ell=1}^{\infty} p_i^\ell (1-p_i)^2} = \frac{1}{\sum_{i=1}^{M} p_i(1-p_i)} = \frac{1}{1-\sum_{i=1}^{M} p_i^2}\,. \tag{A2}$$

Analogously we can derive Eq. (8):

$$\langle \ell \rangle_w = \frac{1}{L}\sum_{\ell=1}^{\infty}\sum_{i=1}^{M} \ell^2 N_i^0(\ell) = \sum_{\ell=1}^{\infty}\sum_{i=1}^{M} \ell^2 p_i^\ell (1-p_i)^2 = \sum_{i=1}^{M}(1-p_i)^2 \sum_{\ell=1}^{\infty}\left(\frac{d}{d\ln p_i}\right)^2 p_i^\ell\,. \tag{A3}$$

Thus,

$$\langle \ell \rangle_w = \sum_{i=1}^{M}(1-p_i)^2\left(p_i\frac{d}{dp_i}\right)^2 \sum_{\ell=1}^{\infty} p_i^\ell = \sum_{i=1}^{M}(1-p_i)^2\left(p_i\frac{d}{dp_i}\right)^2 \frac{p_i}{1-p_i} = 1 + 2\sum_{i=1}^{M}\frac{p_i^2}{1-p_i}\,. \tag{A4}$$

APPENDIX B: DISPERSION OF R_{N} AND R_{W} DUE TO THE FINITE LENGTH OF THE SEQUENCE

Let us denote the probability density of finding an oligomer of length ℓ by $\mathcal{P}_i(\ell)$, where $i = 1, \ldots, M$. For the uncorrelated random sequence,

$$\mathcal{P}_i(\ell) = \frac{p_i^\ell (1-p_i)^2}{1-\chi}\,, \tag{B1}$$

where $\chi = \sum_{i=1}^{M} p_i^2$. The dispersion in the oligomer length is

$$\sigma^2(\ell) = \langle \ell^2 \rangle - \langle \ell \rangle^2 = \frac{1}{1-\chi}\left[1 + 2\sum_{i=1}^{M}\frac{p_i^2}{1-p_i}\right] - \left(\frac{1}{1-\chi}\right)^2 \approx \frac{\chi(1-2\chi)}{(1-\chi)^2} + \mathcal{O}(p_i^3)\,. \tag{B2}$$

The dispersion of the average value of length $\sigma^2(\langle \ell \rangle)$ due to the finite size of the system is

$$\sigma^2(\langle \ell \rangle) = N\sigma^2(\ell)\,. \tag{B3}$$

Since $N = L(1-\chi)$ (see Eqs. (4) and (7)), we find that using Eq. (B2)

$$\sigma(\langle \ell \rangle) \approx \frac{1}{\sqrt{L}}\frac{\sqrt{\chi(1-2\chi)}}{(1-\chi)^2} + \mathcal{O}(p_i^3)\,. \tag{B4}$$

The maximal value of the $\max \sigma(\langle \ell \rangle)\sqrt{L} \approx 0.79$ is achieved when $\chi \approx 0.39$. This is consistent with our numerical analysis (not shown).

The standard deviation of the R_n values is related to the standard deviation of the average oligomer length:

$$\sigma^2(R_n) = \frac{\sigma^2(\langle\ell\rangle)}{\langle\ell\rangle_n^{\rm th} - 1} = \sigma^2(\langle\ell\rangle)\frac{1-\chi}{\chi}\,. \tag{B5}$$

Thus,

$$\sigma(R_n) \approx \frac{1}{\sqrt{L}}\sqrt{\frac{\chi(1-2\chi)}{(1-\chi)^3}} + \mathcal{O}(p_i^3)\,, \tag{B6}$$

so $\max\sigma(R_n)\sqrt{L} \approx 1.19$ when $\chi = 1/4$.

Analogously, we calculate $\sigma(R_w)$:

$$\sigma(R_w) = \frac{\sigma(\langle\ell\rangle_w)}{\sqrt{\langle\ell\rangle_w - 1}} \approx \frac{2.2}{\sqrt{L}} + \mathcal{O}(p_i^3)\,. \tag{B7}$$

APPENDIX C: RATIO MEASURES FOR MARKOV SEQUENCES

For the random Markov sequence, defined by a $M \times M$ matrix $||\Pi_{ij}||$, whose elements Π_{ij} are probabilities of finding element i after element j, the probability density $\mathcal{P}_i(\ell)$ of finding an oligomer of length ℓ is [6]

$$\mathcal{P}_i(\ell) = \frac{p_i\Pi_{ii}^{\ell-1}(1-\Pi_{ii})^2}{1-\chi_M}\,, \tag{C1}$$

where $\chi_M = \sum_{i=1}^{M} p_i\Pi_{ii}$.

The average length of oligomers $\langle\ell\rangle_{n,M}$ is computed in analogy to Appendix A:

$$\langle\ell\rangle_{n,M} = \frac{1}{1-\chi_M}\,, \tag{C2}$$

and

$$\langle\ell\rangle_{w,M} \equiv \langle\ell^2\rangle = \frac{2\chi_M + 1}{1-\chi_M} + \mathcal{O}(p_i\Pi_{ii}^2)\,. \tag{C3}$$

Following the arguments of Appendix B, we find

$$\sigma_M(R_n) \approx \frac{1}{\sqrt{L}}\sqrt{1-2\chi_M} + \mathcal{O}(p_i\Pi_{ii}^2)\,, \tag{C4}$$

and

$$\sigma_M(R_w) \approx \frac{1}{\sqrt{L}}\sqrt{\frac{6-7\chi_M}{3(1-\chi_M)^2}} + \mathcal{O}(p_i\Pi_{ii}^2)\,. \tag{C5}$$

Thus we see that the errors in ratio values due to the finite length L ($> 10^5$) of the Markov sequences are negligible.

[1] 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.

[2] Gacy, A.M., Goellner, G., Juramic, N., Macura, S. and McMurray, C.T., (1995) Trinucleotide repeats that expand in human disease form hairpin structures in vitro. *Cell*, **81**, 553–540.

[3] Richards, R.I. and Sutherland, G.R. (1992) Heritable unstable DNA sequences. *Nature Genetics*, **1**, 7–9.

[4] La Spada, A.R., Paulson, H.L. and Fischbeck, K.H. (1994) Trinucleotide repeat expansion in neurological disease. *Ann. Neurol.*, **36**, 814–822.

[5] Sutherland, G.R. and Richards, R.I. (1995) Simple tandem DNA repeats and human genetic disease. *P.N.A.S. USA*, **92**, 3636–3641.

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

[7] Dokholyan, N. V., Buldyrev, S. V., Havlin, S., and Stanley, H. E. (1998) Non-randomness of dimeric tandem repeats in noncoding DNA sequences. *J. Theor. Biol.* submitted.

[8] Weber, J.L. and May, P.E. (1989) Abundant class of human DNA polymorphisms which can be typed using the Polymerase Chain Reaction. *American Journal of Human Genetics*, **44**, 388–396.

[9] Silver, L.M. (1992) Bouncing off microsatellites. *Nature Genetics*, **2**, 8–9.

[10] Rubinsztein, D.C., Amos, W., Leggo, J., Goodburn, S., Jain, S., Li, S.H., Margolis, R.L., Ross, C.A., Ferguson-Smith, M.A. (1995) Microsatellite evolution: evidence for directionality and variation in rate between species. *Nature Genetics*, **10**, 337–343.

[11] Jin, L. and Chakraborty, R. (1994) Population dynamics of DNA fingerprint patterns within and between populations. *Genet. Res.*, **63**, 1–9.

[12] Nussinov, R. (1980) Strong adenine clustering in nucleotide sequences. *J. Theor. Biol.*, **85**, 285–291.

[13] Sprizhitsky, Y.A., Nechipurenko, Y.D., Alexandrov, A.A. and Volkenstein, M.V. (1988) Statistical analysis of nucleotide runs in coding and noncoding DNA sequences. *Journal of Bio. Struct. and Dynamics*, **6**, 345–358.

[14] Nussinov, R. (1991) Compositional variations in DNA sequences. *CABIOS*, **7**, 287–293.

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

[16] Nussinov, R. (1984) Doublet frequencies in evolutionary distinct groups. *Nucleic Acids Research*, **12**, 1749–1763.

[17] Mrazek, J. and Kypr, K. (1994) Biased distribution of adenine and thymine in gene nucleotide sequences. *Journal of Mol Evol.*, **39**, 439–447.

[18] Borštnik, B., Pumpernik, D., Lukman, D., Ugarković, D., and Plohl, M. (1994) Tandemly repeated pentanucleotides in DNA sequences of eucaryotes. *Nucl. Acid Res.*, **22**, 3412–3417.

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

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

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

[22] Bunde, A. and Havlin, S. (1991) *Fractals and disordered systems.* Springer-Verlag, Berlin.

[23] Stauffer, D. and Aharony, A. (1992) *Introduction to percolation theory.* Taylor & Francis, Philadelphia.

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

[25] Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. (1989) *Numerical Recepies.* Cambridge University Press, Cambridge.

[26] Bell, G.I. (1996) Evolution of simple sequence repeats. *Comp. & Chem.*, **20**, 41–48.

[27] Dokholyan, N. V., Buldyrev, S. V., Havlin, S., and Stanley, H. E. (1998) Model of unequal chromosomal crossing over in DNA sequences. *Physica* **A249**, 594-599.

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

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

[30] Herzel, H., Trifonov, E. N., Weiss, O. and Große, I. (1998) Interpreting correlations in biosequences. *Physica* **A248**, 449–459.

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

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

[33] Dokholyan, N. V., Buldyrev, S. V., Stanley, H. E., and Shakhnovich, E. I. (1998) Molecular dynamics studies of folding of a protein-like model. *Folding & Design* **3,** 577-587.

Table I. The total length in bp of the coding and noncoding regions analyzed. 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

Table II. 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 versus noncoding DNA is less pronounced.

Monomers						
organism	R_n			R_w		
	coding	noncoding	p-value	coding	noncoding	p-value
Primates	1.09 ± 0.01	1.26 ± 0.01	$< 2 \cdot 10^{-5}$	1.08 ± 0.01	1.43 ± 0.01	$< 2 \cdot 10^{-5}$
Vertebrates	1.03 ± 0.01	1.14 ± 0.01	$< 2 \cdot 10^{-5}$	1.02 ± 0.01	1.24 ± 0.01	$< 2 \cdot 10^{-5}$
Invertebrates	1.09 ± 0.01	1.40 ± 0.01	$< 2 \cdot 10^{-5}$	1.08 ± 0.01	1.58 ± 0.01	$< 2 \cdot 10^{-5}$
Mammals	1.06 ± 0.01	1.28 ± 0.02	$< 2 \cdot 10^{-5}$	1.04 ± 0.02	1.43 ± 0.04	$< 2 \cdot 10^{-5}$
Rodents	1.06 ± 0.01	1.18 ± 0.01	$< 2 \cdot 10^{-5}$	1.03 ± 0.01	1.30 ± 0.01	$< 2 \cdot 10^{-5}$
Plants	1.13 ± 0.01	1.10 ± 0.01	$< 2 \cdot 10^{-5}$	1.12 ± 0.01	1.17 ± 0.01	$< 2 \cdot 10^{-5}$
Dimers						
	R_n			R_w		
	coding	noncoding	p-value	coding	noncoding	p-value
Primates	0.96 ± 0.01	1.39 ± 0.01	$< 2 \cdot 10^{-5}$	0.95 ± 0.01	1.73 ± 0.01	$< 2 \cdot 10^{-5}$
Vertebrates	1.00 ± 0.02	1.32 ± 0.02	$< 2 \cdot 10^{-5}$	1.00 ± 0.02	1.81 ± 0.13	$< 2 \cdot 10^{-5}$
Invertebrates	0.95 ± 0.01	1.39 ± 0.01	$< 2 \cdot 10^{-5}$	0.97 ± 0.01	1.49 ± 0.01	$< 2 \cdot 10^{-5}$
Mammals	0.94 ± 0.02	1.43 ± 0.04	$< 2 \cdot 10^{-5}$	0.94 ± 0.02	1.74 ± 0.09	$< 2 \cdot 10^{-5}$
Rodents	0.93 ± 0.01	1.47 ± 0.01	$< 2 \cdot 10^{-5}$	0.93 ± 0.01	2.33 ± 0.01	$< 2 \cdot 10^{-5}$
Plants	0.95 ± 0.01	1.21 ± 0.01	$< 2 \cdot 10^{-5}$	0.95 ± 0.01	1.36 ± 0.03	$< 2 \cdot 10^{-5}$

Trimers						
	R_n			R_w		
	coding	noncoding	p-value	coding	noncoding	p-value
Primates	1.51 ± 0.02	1.49 ± 0.01	0.31	1.63 ± 0.03	1.91 ± 0.02	$1.7 \cdot 10^{-4}$
Vertebrates	1.54 ± 0.05	1.40 ± 0.04	$1.7 \cdot 10^{-4}$	1.68 ± 0.08	1.56 ± 0.06	$3.1 \cdot 10^{-2}$
Invertebrates	1.49 ± 0.01	1.21 ± 0.01	$1.7 \cdot 10^{-4}$	1.56 ± 0.01	1.27 ± 0.01	$1.7 \cdot 10^{-4}$
Mammals	1.53 ± 0.04	1.51 ± 0.04	0.97	1.63 ± 0.06	1.87 ± 0.10	$1.7 \cdot 10^{-4}$
Rodents	1.42 ± 0.02	1.40 ± 0.01	$6.9 \cdot 10^{-3}$	1.52 ± 0.02	2.13 ± 0.07	$< 2 \cdot 10^{-5}$
Plants	1.42 ± 0.01	1.29 ± 0.02	$1.7 \cdot 10^{-4}$	1.50 ± 0.01	1.45 ± 0.02	$1.7 \cdot 10^{-4}$

Tetramers						
	R_n			R_w		
	coding	noncoding	p-value	coding	noncoding	p-value
Primates	0.85 ± 0.02	2.85 ± 0.03	$< 2 \cdot 10^{-5}$	0.86 ± 0.02	4.61 ± 0.07	$< 2 \cdot 10^{-5}$
Vertebrates	0.89 ± 0.04	2.57 ± 0.19	$< 2 \cdot 10^{-5}$	0.89 ± 0.04	5.71 ± 1.17	$< 2 \cdot 10^{-5}$
Invertebrates	0.83 ± 0.01	1.31 ± 0.01	$< 2 \cdot 10^{-5}$	0.96 ± 0.02	1.53 ± 0.02	$< 2 \cdot 10^{-5}$
Mammals	0.68 ± 0.03	2.96 ± 0.29	$< 2 \cdot 10^{-5}$	0.69 ± 0.03	3.84 ± 0.46	$< 2 \cdot 10^{-5}$
Rodents	0.79 ± 0.02	4.57 ± 0.06	$< 2 \cdot 10^{-5}$	0.80 ± 0.02	11.32 ± 0.23	$< 2 \cdot 10^{-5}$
Plants	0.91 ± 0.01	1.85 ± 0.03	$< 2 \cdot 10^{-5}$	0.92 ± 0.01	2.27 ± 0.15	$< 2 \cdot 10^{-5}$

Table III. The average clustering ratio values R_n and R_w along with the error bars are shown for 16 dimers in the primate taxonomic partition of the GenBank for coding and noncoding DNA. Note that the standard deviation of the clustering measures is a few per cent, indicating conservation of the clustering measures for each particular dimer within the same organism.

Dimers: Primates				
dimer	R_n		R_w	
	coding	noncoding	coding	noncoding
AA	1.36 ± 0.04	2.07 ± 0.04	1.35 ± 0.04	2.92 ± 0.09
AT	0.77 ± 0.02	1.17 ± 0.01	0.78 ± 0.02	1.39 ± 0.02
AG	0.93 ± 0.01	1.13 ± 0.01	0.93 ± 0.01	1.18 ± 0.01
AC	1.05 ± 0.01	1.61 ± 0.02	1.05 ± 0.01	2.11 ± 0.05
TA	1.70 ± 0.03	1.48 ± 0.02	1.70 ± 0.03	1.80 ± 0.04
TT	1.50 ± 0.02	1.95 ± 0.03	1.50 ± 0.03	2.70 ± 0.07
TG	0.89 ± 0.01	1.04 ± 0.01	0.89 ± 0.02	1.37 ± 0.03
TC	1.31 ± 0.02	1.72 ± 0.01	1.29 ± 0.02	1.78 ± 0.02
GA	1.22 ± 0.02	1.68 ± 0.01	1.20 ± 0.02	1.73 ± 0.01
GT	1.60 ± 0.04	1.55 ± 0.01	1.58 ± 0.04	2.27 ± 0.07
GG	0.80 ± 0.01	1.18 ± 0.01	0.77 ± 0.01	1.15 ± 0.01
GC	0.46 ± 0.01	0.34 ± 0.01	0.47 ± 0.01	0.35 ± 0.01
CA	0.71 ± 0.02	1.05 ± 0.01	0.72 ± 0.02	1.34 ± 0.03
CT	0.81 ± 0.01	1.14 ± 0.01	0.80 ± 0.01	1.20 ± 0.02
CG	1.25 ± 0.03	2.05 ± 0.06	1.26 ± 0.03	2.11 ± 0.06
CC	0.95 ± 0.01	1.18 ± 0.01	0.91 ± 0.01	1.15 ± 0.01