*Journal of Biomolecular Structure & Dynamics, ISSN 0739-1102*
*Volume 17, Issue Number 1, (1999)*


# Clustering of Identical Oligomers in Coding and Noncoding DNA Sequences

http://www.adeninepress.com

**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

***Abstract***

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.

## *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 (8). Since microsatellites were first shown to aid genetic mapping (9), they have become primary genetic markers (10). More recently, studies have used microsatellites to compare evolution among different species (11). At some loci, microsatellites are so polymorphic that they can be used for DNA fingerprinting (12).

Quantitative studies of microsatellites include analysis of runs of single nucleotides (13,14). Dinucleotides have been studied in terms of nearest neighbors (15,16) and relative frequencies (6,7,17). Trinucleotides have been examined in terms of biased distributions (18). Tandemly repeated pentamers have been studied in (19). A comprehensive study of the average length of simple repeats of units of 1—6 nucleotides was compiled (20). An analysis of clustering of nucleotides has been done by Mrazek and Kypr (21) and by Lio et al. for *Haemophilus influenzae* and *Saccharomyces cerevisiae* chromosomes (22).

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 a difference in the evolutionary pressure in various functional parts of DNA. Recent studies of distri-

*Phone: 617-353-8936;
Fax: 617-353-9393;
E-mail: dokh@bu.edu

butions 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. Using the concepts of percolation theory (23-25), we calculate the mean length (defined below) of repetitions of oligomers, and the expected length of repetitions if the oligomers, with the same frequencies as in a real sequence, were randomly placed along an artificial sequence. The expected length of repetitions of oligomers 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 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.

We systematically compare clustering in coding and noncoding DNA for: 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.

***Method: Ratio Analysis***

We quantify the repetitions of oligomers by dividing the sequence into the nonoverlapping 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 frames.

We analyze separately coding and noncoding sequences. For coding sequences we concatenate exons within a single gene (excluding the untranslated 5′ and 3′ ends[1]). 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 and the number of sequences analyzed in coding and noncoding regions of different taxonomic partitions are reported in Table I.

**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 | non-coding | 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 |

First, we compute the number of repeats of length $\ell$ of a given oligomer in the analyzed set of sequences: $N_i(\ell)$, 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

$$\sum_{\ell=1}^{\infty} \sum_{i=1}^{M} N_i(\ell)\ell = L\,, \qquad [1]$$

where L is the total number of oligomers.

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

[1]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.

$$\langle \ell \rangle_n \equiv \frac{L}{N}, \qquad [2]$$

where

$$N = \sum_{\ell=1}^{\infty} \sum_{i=1}^{M} N_i(\ell) \qquad [3]$$

is the total number of repeat occurrences.

(*ii*) We also define the "weight" average (see e. g. [25]):

$$\langle \ell \rangle_w \equiv \frac{\sum_{\ell=1}^{\infty} \sum_{i=1}^{M} \ell^2 N_i(\ell)}{\sum_{\ell=1}^{\infty} \sum_{i=1}^{M} \ell N_i(\ell)} = \frac{\sum_{\ell=1}^{\infty} \sum_{i=1}^{M} \ell^2 N_i(\ell)}{L}. \qquad [4]$$

This definition gives larger weights to longer repeats. The utility of [4] 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, we can compute the probability $P_\ell$ that a given oligomer belongs to a "cluster" (aggregate) of *exactly* $\ell$ repetitions in an uncorrelated random sequence[2] is given by percolation theory (23,24):

$$P_i(\ell) = \ell p^{\ell}{}_i \ (1 - p_i)^2 \qquad [5]$$

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

$$N_i^0(\ell) = L p^{\ell}{}_i \ (1 - p_i)^2. \qquad [6]$$

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

$$\langle \ell \rangle_n^{\text{th}} = \frac{1}{1 - \sum_{i=1}^{M} p_i^2}, \qquad [7]$$

where $p_i$ is the frequency of each oligomer, and

$$\langle \ell \rangle_w^{\text{th}} = 1 + 2 \sum_{i=1}^{M} \frac{p_i^2}{1 - p_i}. \qquad [8]$$

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

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

The clustering ratios compare the actual repeat length with the control case in

[2]For the Markov sequence consult Appendix C.

which, by definition, no clustering occurs beyond the clustering that occurs in an 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. Their standard deviations are computed in Appendix B. In Table II, we present the relative clustering ratios $R_n$ and $R_w$.

### *Results and Discussion*

We compare the ratios of the observed values of the average length $\langle \ell \rangle_n$ of oligomers[3] (monomers, dimers, trimers, and tetramers) and their weight average

**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. 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*) ≈ 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.

| Monomers | | | | | | |
|---|---|---|---|---|---|---|
| Organism | $R_n$ | | | $R_w$ | | |
| | coding | non-coding | *p*-value | coding | non-coding | *p* -value |
| Primates | 1.09 ± 0.01 | 1.26 ± 0.01 | $<2\cdot10^{-5}$ | 1.08 ± 0.01 | 1.43 ± 0.01 | $<2\cdot10^{-5}$ |
| Vertebrates | 1.03 ± 0.01 | 1.14 ± 0.01 | $<2\cdot10^{-5}$ | 1.02 ± 0.01 | 1.24 ± 0.01 | $<2\cdot10^{-5}$ |
| Invertebrates | 1.09 ± 0.01 | 1.40 ± 0.01 | $<2\cdot10^{-5}$ | 1.08 ± 0.01 | 1.58 ± 0.01 | $<2\cdot10^{-5}$ |
| Mammals | 1.06 ± 0.01 | 1.28 ± 0.02 | $<2\cdot10^{-5}$ | 1.04 ± 0.02 | 1.43 ± 0.04 | $<2\cdot10^{-5}$ |
| Rodents | 1.06 ± 0.01 | 1.18 ± 0.01 | $<2\cdot10^{-5}$ | 1.03 ± 0.01 | 1.30 ± 0.01 | $<2\cdot10^{-5}$ |
| Plants | 1.13 ± 0.01 | 1.10 ± 0.01 | $<2\cdot10^{-5}$ | 1.12 ± 0.01 | 1.17 ± 0.01 | $<2\cdot10^{-5}$ |
| Dimers | | | | | | |
| | $R_n$ | | | $R_w$ | | |
| | coding | non-coding | *p* -value | coding | non-coding | *p* -value |
| Primates | 0.96 ± 0.01 | 1.39 ± 0.01 | $<2\cdot10^{-5}$ | 0.95 ± 0.01 | 1.73 ± 0.01 | $<2\cdot10^{-5}$ |
| Vertebrates | 1.00 ± 0.02 | 1.32 ± 0.02 | $<2\cdot10^{-5}$ | 1.00 ± 0.02 | 1.81 ± 0.13 | $<2\cdot10^{-5}$ |
| Invertebrates | 0.95 ± 0.01 | 1.39 ± 0.01 | $<2\cdot10^{-5}$ | 0.97 ± 0.01 | 1.49 ± 0.01 | $<2\cdot10^{-5}$ |
| Mammals | 0.94 ± 0.02 | 1.43 ± 0.04 | $<2\cdot10^{-5}$ | 0.94 ± 0.02 | 1.74 ± 0.09 | $<2\cdot10^{-5}$ |
| Rodents | 0.93 ± 0.01 | 1.47 ± 0.01 | $<2\cdot10^{-5}$ | 0.93 ± 0.01 | 2.33 ± 0.01 | $<2\cdot10^{-5}$ |
| Plants | 0.95 ± 0.01 | 1.21 ± 0.01 | $<2\cdot10^{-5}$ | 0.95 ± 0.01 | 1.36 ± 0.03 | $<2\cdot10^{-5}$ |
| Trimers | | | | | | |
| | $R_n$ | | | $R_w$ | | |
| | coding | non-coding | *p* -value | coding | non-coding | *p* -value |
| Primates | 1.51 ± 0.02 | 1.49 ± 0.01 | 0.31 | 1.63 ± 0.03 | 1.91 ± 0.02 | $1.7\cdot10^{-4}$ |
| Vertebrates | 1.54 ± 0.05 | 1.40 ± 0.04 | $1.7\cdot10^{-4}$ | 1.68 ± 0.08 | 1.56 ± 0.06 | $3.1\cdot10^{-2}$ |
| Invertebrates | 1.49 ± 0.01 | 1.21 ± 0.01 | $1.7\cdot10^{-4}$ | 1.56 ± 0.01 | 1.27 ± 0.01 | $1.7\cdot10^{-4}$ |
| Mammals | 1.53 ± 0.04 | 1.51 ± 0.04 | 0.97 | 1.63 ± 0.06 | 1.87 ± 0.10 | $1.7\cdot10^{-4}$ |
| Rodents | 1.42 ± 0.02 | 1.40 ± 0.01 | $6.9\cdot10^{-3}$ | 1.52 ± 0.02 | 2.13 ± 0.07 | $<2\cdot10^{-5}$ |
| Plants | 1.42 ± 0.01 | 1.29 ± 0.02 | $1.7\cdot10^{-4}$ | 1.50 ± 0.01 | 1.45 ± 0.02 | $1.7\cdot10^{-4}$ |
| Tetramers | | | | | | |
| | $R_n$ | | | $R_w$ | | |
| | coding | non-coding | *p* -value | coding | non-coding | *p* -value |
| Primates | 0.85 ± 0.02 | 2.85 ± 0.03 | $<2\cdot10^{-5}$ | 0.86 ± 0.02 | 4.61 ± 0.07 | $<2\cdot10^{-5}$ |
| Vertebrates | 0.89 ± 0.04 | 2.57 ± 0.19 | $<2\cdot10^{-5}$ | 0.89 ± 0.04 | 5.71 ± 1.17 | $<2\cdot10^{-5}$ |
| Invertebrates | 0.83 ± 0.01 | 1.31 ± 0.01 | $<2\cdot10^{-5}$ | 0.96 ± 0.02 | 1.53 ± 0.02 | $<2\cdot10^{-5}$ |
| Mammals | 0.68 ± 0.04 | 2.96 ± 0.29 | $<2\cdot10^{-5}$ | 0.69 ± 0.04 | 3.84 ± 0.46 | $<2\cdot10^{-5}$ |
| Rodents | 0.79 ± 0.02 | 4.57 ± 0.06 | $<2\cdot10^{-5}$ | 0.80 ± 0.02 | 11.32 ± 0.23 | $<2\cdot10^{-5}$ |
| Plants | 0.91 ± 0.01 | 1.85 ± 0.03 | $<2\cdot10^{-5}$ | 0.92 ±0.01 | 2.27 ± 0.15 | $<2\cdot10^{-5}$ |

[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.

$\langle \ell \rangle_w$ 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[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, each 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 (26)). If *p* is close to 1, then the two sets of values are drawn from the same distribution with the probability 1. If *p* is close to 0, then these set of values are taken from two different distributions with the probability $(1\text{-}p) \rightarrow 1$. 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.

(*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. In contrast, 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. 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 II), which indicates that one cannot distinguish between coding and noncoding DNA based only on $R_n$ 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.

Observations (*i*) - (*iii*) might arise from the evolutionary pressure against clustering of repeats (except trimeric) in coding DNA. The source of clustering of oligomers in noncoding DNA could be the result of various duplication processes or simple repeat expansion processes (5,6,27), indicating that some of the neighboring oligomers evolved from the same single copy. These observations are in agreement with recent work (7,16,28), where dimeric tandem repeats (DTR) were studied and it was found that DTR are abundant in noncoding DNA, while they are rare in coding DNA. The difference in length distributions of DTR 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. (22).

[4]We studied the complete genome of *Escherichia coli*; however we found that clustering ratios are close to unity both in coding and noncoding DNA.

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 the native state of the folded proteins (29-32).

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 (33-34)), would not affect the protein function, and hence would be tolerated by natural selection.

We also calculate the clustering measures for each individual oligomer, which we define the same way as in Eqs. (2) - (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

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

and

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

The theoretical values for $\langle \ell \rangle_{n,i}^{\rm th} = 1/(1\text{-}p_i)$ and $\langle \ell \rangle_{w,i}^{\rm th} = (1 + p_i)/(1\text{-}p_i)$ (see (25)) are computed similarly to Eq. (7) and (8).

We find that the clustering ratios for each individual oligomer are conserved for each organism, i.e. the standard deviation of the clustering measures in various parts of the same genome is around a few percent. To illustrate this observation we report the clustering ratio values for dimers in *Homo sapiens* in Table III[5]. This

**Table III**

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

| Dimers: *Homo sapiens* | | | | |
|---|---|---|---|---|
| dimer | $R_n$ | | $R_w$ | |
| | coding | non-coding | coding | non-coding |
| 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 |

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

observation indicates that the clustering ratio can quantify the tendency 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.

***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 of the length average $\langle \ell \rangle_n$:

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

where $N_i^0(\ell)$, the number of oligomers of type $i$ in a sequence of length $L$, is determined by the Eq. (6). Thus,

$$\langle \ell \rangle_n = \frac{1}{\sum_{i=1}^{M} \sum_{\ell=1}^{\infty} p_i^{\ell}(1-p_i)^2} = \frac{1}{\sum_{i=1}^{M} p_i(1-p_i)} = \frac{1}{1-\sum_{i=1}^{M} p_i^2}. \qquad \text{[A2]}$$

Analogously we can derive Eq. (8):

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

Thus,

$$\langle \ell \rangle_w = \sum_{i=1}^{M} (1-p_i)^2 \left(p_i \frac{d}{dp_i}\right)^2 \sum_{\ell=1}^{\infty} p_i^{\ell} = \sum_{i=1}^{M} (1-p_i)^2 \left(p_i \frac{d}{dp_i}\right)^2 \frac{p_i}{1-p_i} = 1 + 2\sum_{i=1}^{M} \frac{p_i^2}{1-p_i} \qquad \text{[A4]}$$

***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 a cluster of length $\ell$ by $P_i(\ell)$, where $i$ =1, ..., $M$. For the uncorrelated random sequence,

$$\mathcal{P}_i(\ell) \equiv \frac{N_i^0(\ell)}{N} = \frac{p_i^{\ell}(1-p_i)^2}{1-\chi}, \qquad \text{[B1]}$$

where $\chi = \sum_{i=1}^{M} p_i^2$ $\chi$ = i=1M pi2. The dispersion of the cluster length is

$$\sigma^2(\ell) = \langle \ell^2 \rangle - \langle \ell \rangle^2 = \frac{1}{1-\chi}\left[1 + 2\sum_{i=1}^{M} \frac{p_i^2}{1-p_i}\right] - \left(\frac{1}{1-\chi}\right)^2 = \chi + o(\chi). \qquad \text{[B2]}$$

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

$$\sigma^2(\langle \ell \rangle) = \frac{1}{N}\sigma^2(\ell). \qquad \text{[B3]}$$

Since $N = L(1-\chi)$ (see Eqs. (2) and (7)), we find using Eq. (2) that

$$\sigma(\langle \ell \rangle) = \frac{1}{\sqrt{L}}(\sqrt{\chi} + o(\sqrt{\chi})). \qquad \text{[B4]}$$

Since the dispersion due to the errors in measuring $p_i$ is negligible, the standard deviation of the $R_n$ values can be computed as follows:

$$\sigma^2(R_n) = \frac{\sigma^2(\langle \ell \rangle)}{(\langle \ell \rangle_n^{\text{th}} - 1)^2} \approx \sigma^2(\langle \ell \rangle)\frac{1}{\chi^2}. \quad \text{[B5]}$$

Thus,

$$\sigma(R_n) \approx \frac{1}{\sqrt{L}}\frac{1}{\sqrt{\chi}}. \quad \text{[B6]}$$

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

$$\sigma(R_w) = \frac{\sigma(\langle \ell \rangle_w)}{\langle \ell \rangle_w - 1} \approx \frac{1}{\sqrt{L}}\frac{1}{\sqrt{\chi}}. \quad \text{[B7]}$$

According to the Cauchy-Bounyakovsky inequality, the minimal value of $\chi$ is 1/*M*. Using Eqs. [B6] and [B7], we estimate the dispersion due to the finite sampling size for all entries of Table I. In all cases the expected fluctuations of clustering ratios due to the finite sampling size do not exceed the error found directly by analyzing different data segments and, hence, can be neglected.

### *Appendix C: Ratio Measures for Markov Sequences*

For the one-step Markov sequence, generated with the help of a $M \times M$ matrix $\|\Pi_{ij}\|$, whose elements $\Pi_{ij}$ are probabilities of finding element *i* after element *j*, the probability density $P_i(\ell)$ of finding an oligomer of length $\ell$ is (6)

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

where $\chi_M = \Sigma_{i=1}^{M} p_i \Pi_{ij}$.

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

$$\langle \ell \rangle_{n,M} = \frac{1}{1 - \chi_M}, \quad \text{[C2]}$$

and analogously

$$\langle \ell \rangle_{w,M} = (1 + 2\chi_M). \quad \text{[C3]}$$

Following the arguments of Appendix B, we find

$$\sigma_M(R_n) \approx \frac{1}{\sqrt{L}}\frac{1}{\sqrt{\chi_M}}, \quad \text{[C4]}$$

and

$$\sigma_M(R_w) \approx \frac{1}{\sqrt{L}}\frac{1}{\sqrt{\chi_M}}. \quad \text{[C5]}$$

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

#### *References and Footnotes*

1. Huntington's Disease Collaborative Research Group, *Cell 72*, 971-983 (1993).
2. Gacy, A.M., Goellner, G., Juramic, N., Macura, S. and McMurray, C.T., *Cell 81*, 553-540

(1995).
3. Richards, R.I. and Sutherland, G.R., *Nature Genetics 1*, 7-9 (1992).
4. La Spada, A.R., Paulson, H.L. and Fischbeck, K.H., *Ann. Neurol. 36*, 814-822 (1992).
5. Sutherland, G.R. and Richards, R.I., *P.N.A.S. USA 92*, 3636-3641 (1995).
6. Dokholyan, N.V., Buldyrev, S.V., Havlin, S. and Stanley, H.E., *Phys. Rev. Lett. 79*, 5182-5185 (1997).
7. Dokholyan, N.V., Buldyrev, S.V., Havlin, S. and Stanley, H.E., *J. Theor. Biol.* submitted (1998).
8. Burge, C., Campbell, A.M. and Karlin, S., *Proc. Natl. Acad. Sci. USA 89*, 1358-1362 (1992).
9. Weber, J.L. and May, P.E., *American Journal of Human Genetics 4*, 388-396 (1989).
10. Silver, L.M., *Nature Genetics 2*, 8-9 (1992).
11. Rubinsztein, D.C., Amos, W., Leggo, J., Goodburn, S., Jain, S., Li, S.H., Margolis, R.L., Ross, C.A. and Ferguson-Smith, M.A., *Nature Genetics 10*, 337-343 (1995).
12. Jin, L. and Chakraborty, R., *Genet. Res. 63*, 1-9 (1994).
13. Nussinov, R., *J. Theor. Biol. 85*, 285-291 (1980).
14. Sprizhitsky, Y.A., Nechipurenko, Y.D., Alexandrov, A.A. and Volkenstein, M.V., *Journal of Struct. and Dynamics 6*, 345-358 (1988).
15. Nussinov, R., *CABIOS 7*, 287-293 (1991).
16. Bell, G.I. and Jurka, J., *J. Mol. Evol. 44*, 414-421 (1997).
17. Nussinov, R., *Nucleic Acids Research 2*, 1749-1763 (1984).
18. Mrazek, J. and Kypr, K., *Journal of Mol. Evol. 39*, 439-447 (1994).
19. Borstnik, B., Pumpernik, D., Lukman, D., Ugarkovic, D. and Plohl, M., *Nucl. Acid Res. 22*, 3412-3417 (1994).
20. Jurka, J. and Pethiyagoda, C., *J. of Mol. Evol. 40*, 120-126 (1995).
21. Mrazek, J. and Kypr, K., *CABIOS 11*, 195-199 (1995).
22. Lio, P., Politi, A., Ruffo, S. and Buiatti, M., *J. Theor. Biol. 183*, 455-469 (1996).
23. Bunde, A. and Havlin, S., *Fractals and disordered systems*. Springer-Verlag, Berlin (1991).
24. Stauffer, D. and Aharony, A., *Introduction to percolation theory*. Taylor & Francis, Philadelphia (1992).
25. Reynolds, P.J., Stanley, H.E. and Klein, W., *J. Phys. A 10*, L203210 (1977).
26. Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T., *Numerical Recepies*. Cambridge University Press, Cambridge (1989).
27. Dokholyan, N.V., Buldyrev, S.V., Havlin, S. and Stanley, H.E., *Physica A249*, 594-599 (1998).
28. Bell, G.I., *Comp. & Chem. 20*, 41-48 (1996).
29. Shakhnovich, E.I. and Gutin, A.M., *Nature 346*, 773-775 (1990).
30. Abkevich, V., Gutin, A.M. and Shakhnovich, E.I., *J. Mol. Biol. 252*, 460-471 (1995).
31. Herzel, H., Trifonov, E.N., Weiss, O. and Große, I., *Physica A248*,449-459 (1998).
32. Mirny, L.A., Abkevich, V. and Shakhnovich, E.I., *Folding & Design 1*,103-116 (1996).
33. Shakhnovich, E.I., Abkevich, V.I. and Ptitsyn, O., *Nature 379*, 96-98 (1996).
34. Dokholyan, N.V., Buldyrev, S.V., Stanley, H. E. and Shakhnovich,E. I., *Folding & Design 3*, 577-587 (1998).

*Date Received: May 28, 1999*

***Communicated by the Editor Maxim Frank-Kamanetskii***