INFORMATICA, 2010, Vol. 21, No. 1, 13–30

Complexity Estimation of Genetic Sequences Using Information-Theoretic and Frequency Analysis Methods

Robertas DAMAŠEVIČIUS
Software Engineering Department, Kaunas University of Technology
Studentų 50–415, LT-51368 Kaunas, Lithuania
e-mail: robertas.damasevicius@ktu.lt

Received: January 2009; accepted: February 2009

Abstract. The genetic information in cells is stored in DNA sequences, represented by a string of four letters, each corresponding to a definite type of nucleotides. Genomic DNA sequences are very abundant in periodic patterns, which play important biological roles. The complexity of genetic sequences can be estimated using the information-theoretic methods. Low complexity regions are of particular interest to genome researchers, because they indicate to sequence repeats and patterns. In this paper, the complexity of genetic sequences is estimated using Shannon entropy, Rényi entropy and relative Kolmogorov complexity. The structural complexity based on periodicities is analyzed using the autocorrelation function and time delayed mutual information. As a case study, we analyze human 22nd chromosome and identify 3 and 49 bp periodicities.

Keywords: genetic sequence, DNA analysis, entropy, complexity, frequency analysis, bioinformatics.

1. Introduction

The genetic information in cells is stored in DNA sequences, represented by a string of four letters, each corresponding to a definite type of nucleotides: adenine (A), guanine (G), cytosine (C), and thymine (T). These letters can form different combinations. The analysis of genetic sequences is one of the primary tasks of *bioinformatics*. These sequences can be analyzed using applied mathematics, statistical, linguistic, digital signal processing, etc., methods to enable finding and marking the genes and other biological features in a DNA sequence, understand genome evolution, predict gene expression, compare genomes and study evolutionary relationships between organisms.

Although much research has been done in this area, still the function and purpose of much of the DNA sequences is not fully understood. For example, over 95% of DNA of the human genome has been designated as "junk" (Nowak, 1994). While much of this sequence may be an evolutionary artifact that serves no present-day purpose, some is believed to function in ways that are not currently understood. There-

fore, the analysis of structural complexity features of DNA sequences, such as *repeats* (Herzel *et al.*, 1994; Holste *et al.*, 2003), *motifs* (Hancock and Armstrong, 1994), *patterns* (Freeman *et al.*, 1998; Larsabal and Danchin, 2005), *periodicities* (Herzel *et al.*, 1999; Fukushima *et al.*, 2001) and *long-range correlations* (Li and Kaneko, 1992; Mansilla *et al.*, 2004) is considered as an important task of computational genomics.

Complexity of a symbolic sequence reflects an inability to represent a sequence in a compact form based on some structural features of this sequence. To evaluate complexity of DNA sequences, several groups of methods have been developed: *entropy* measures (Schurmann and Grassberger, 1996; Schmitt and Herzel, 1997; Holste *et al.*, 2001; Frappat *et al.*, 2003; Vinga and Almeida, 2004), *Shannon information* and *divergence* (Chang *et al.*, 2005; Chen *et al.*, 2005), *alphabetical symbol frequencies* (Wootton and Federhen, 1996), *evaluation of the n-grams, combinatorial complexity* and *linguistic complexity* (Mantegna *et al.*, 1994; Troyanskaya *et al.*, 2002), modifications of the *Lempel and Ziv complexity* (Chen *et al.*, 1999; Gusev *et al.*, 2003), *Hurst exponent* (Yu and Anh, 2001), *fractal dimension* (Cristea and Popescu, 2003), *stochastic complexity* (Orlov *et al.*, 2002), *grammatical complexity* (Jimenez-Montano *et al.*, 2002), *correlation information* (Hou, 2007) and approaches based on *Kolmogorov complexity* and compression algorithms (Adjeroh *et al.*, 2002).

The other approach to the analysis of the DNA sequences is converting the symbolic DNA sequences into digital signals and applying digital signal processing (DSP) methods such as *Fourier Transform* (Lyshevski and Krueger, 2004), *mutual information functions* (Herzel *et al.*, 1998; Grosse *et al.*, 2000; Holste *et al.*, 2003), *power spectra* (Voss, 1992; Li and Kaneko, 1992; Li, 1997; Fukushima *et al.*, 2002), *auto-correlation analysis* (Azbel, 1995; Herzel *et al.*, 1998; De Sousa Vieira, 1999; Larsabal and Danchin, 2005), *wavelet analysis* (Berger and Mitra, 2002), and *detrended fluctuation analysis* (Peng *et al.*, 1995; Buldyrev *et al.*, 1995). Application of the DSP methods for the analysis of symbolic genomic sequences can reveal the features of genetic sequences that would be difficult to grasp using standard statistical analysis and pattern matching methods.

The aim of this paper is to analyze methods for estimating complexity of genetic sequences and to use these methods to determine the structural features of the genetic sequences. For analysis, the human 22nd chromosome (Dunham *et al.*, 1999; NCBI) was selected. It was the first fully sequenced human chromosome, which is often used as a benchmark in DNA sequence research (see, e.g., Cristea and Popescu, 2003).

The structure of the paper is as follows. Section 2 discusses the complexity of genetic sequences, its definition, related concepts and specific types of complexity in DNA sequences. Section 3 describes the complexity metrics used in this research. Section 4 presents an analysis of the human 22nd chromosome using information-theoretic and frequency analysis methods. Section 5 evaluates the results. Finally, Section 6 presents the conclusions.

2. Complexity of Genetic Sequences

2.1. *What is Complexity?*

Complexity has proven to be a difficult concept to define. Different researchers use very different definitions of complexity, and there are still no single agreed-upon definition. Complexity of a system usually means that we cannot represent it in a short, comprehensive description. Its opposite is *simplicity* – the ability of some general method of analysis to express the content of a system in a short (compressed) form.

In general, the concept of complexity is related to such concepts as order, organization, meaning, randomness, redundancy, compressibility and information. A finite-state sequence is said to be random if all states have equal probabilities at all sequence positions. Thus, the random sequence has no recognizable meaning, structure, patterns or order. Such random sequence is not compressible, which may indicate high complexity. When a sequence contains some repeating patterns, some degree of redundancy emerges, and a sequence becomes more ordered and less complex. Highly-complex systems exhibit a hierarchical internal structure with fractal-like repeating fragments. The existence of repeating fragments allows the sequence to be more compressible. Highly redundant sequence, which contains a great number of simple repeating patterns, is highly ordered, and compressible, but has low complexity. Summarizing, complexity represents a state of an object (e.g., sequence) between order and randomness, which is related to a high number of repeating patterns organized in a complex structure.

The most fundamental type of complexity is *informational* complexity. It is fundamental in the sense that anything that is complex in any way also must be informationally complex. A complex object requires more information to specify it than a simple one. The most distinguishing property of a complex system is its high information content. In that sense, a highly-complex sequence may be undistinguishable from a random sequence: the fact that we cannot recognize the structural organization of a sequence does not mean that there is no such. Thus, the informational complexity of a sequence ultimately depends upon its meaning to the external observer, i.e., the amount of information that can be extracted from the sequence. If the observer cannot extract any information from the sequence, then its information content (and complexity) is low. If the observer recognizes the meaning of the sequence and extracts some useful information, the informational complexity of the sequence may be proportional to the amount of extracted information. In other words, randomness and informational complexity of the sequence is relative to the capacity of the observer to find a "meaning" in the analyzed sequence (Crisan, 2004).

2.2. *Complexity of DNA*

Any DNA sequence looks like a text written in an unknown language, which is encoded in a 4-letter alphabet {A, C, G, T}. Each letter in this text corresponds to a DNA base pair: A – adenine, C – cytosine, G – guanine, and T – thymine. Despite the simplicity of its alphabet, DNA exhibits a great degree of complexity.

Genomic DNA sequence is very abundant in periodic patterns (e.g., isochores, intergenic sequences, CpG islands, LINE (long interspersed elements) and SINE (short interspersed elements) repeats, genes, exons, introns, tandem repeats, variable number tandem repeats, *Alu* repeats, bend sites, satellite DNA, interstitial telomeric sequences, and long-range correlations), which play important biological roles, such as gene expression, genome structural stabilization, and recombination (Buldyrev *et al.*, 1995). A *tandem repeat* is a type of periodic patterns, when a pattern of two or more nucleotides is repeated and the repetitions are directly adjacent to each other. Tandem repeats are related to several genetic diseases. A *variable number tandem repeat* is a short nucleotide sequence ranging from 14 to 100 nucleotides long that is organized into clusters of tandem repeats, usually repeated in the range of between 4 and 40 times per occurrence. Each of these structural elements has its different size distribution, nucleotide frequencies, and laws of molecular evolution, so the correlations in the DNA sequence have very complex structure, and are different for different species.

The analysis of statistical patterns in DNA sequences is important, since correlations may reflect biologically significant features of primary structures. Among the short-range periodicities that have strong biological significance the most important one is $f^{-1} = 3$ bp (bp = base pairs), which is easily found in the genomes of all prokaryotes and lower eukaryotes, but appears very dimly in the genomes of higher eukaryotes due to low gene density in their genomes. This periodicity indicates the presence of *protein-coding* genes in genomes, and can be used to distinguish between the coding and non-coding DNA sequences. The other important periodicity has a mean value of $f^{-1} = 10.5$ bp and reflects *DNA bendability* and *secondary structure* of proteins (Fukushima *et al.*, 2002). Periodicities of $f^{-1} = 102$ bp can be explained by the nucleosomal structure in eukaryotes. Periodicities of about $f^{-1} = 106$ bp are a well-known biological phenomenon related to the presence of *isochores* (regions of DNA with relatively high G + C content; Buldyrev *et al.*, 1995). Next significant periodicities at about $f^{-1} = 135$ bp, $f^{-1} = 165$ bp and $f^{-1} = 300$ bp are related to *Alu* sequences, which do not contain any coding sequences and can be recognized by the restriction enzyme AluI (Holste *et al.*, 2003).

The existence of structures at very large scales results in long-range correlations. *Long-range correlations* extend over distances of hundreds of thousands to tens of millions of bp, i.e., up to the scale of whole chromosomes, and have a functional role in the control of crossing-over and species separation (Peng *et al.*, 1995). They also could represent a trade-off between efficient information storage and protection against error in the genetic code by adding some redundancy to the encoding. Long-range correlations lack universality, i.e., they are different for different species. The periodicities of DNA sequences are summarized in Table 1.

Genomic sequences contain numerous "layers" of information (Stanley *et al.*, 1999). These include specifications for mRNA sequences responsible for protein structure, identification of coding and non-coding parts of the sequence, information necessary for specification of regulatory (promoter, enhancer) sequences, information directing protein-DNA interactions, directions for DNA packaging and unwinding. The genomic sequence

Table 1
Periodicities in DNA sequences

Period (bp)	Meaning	References
3	Protein-coding genes	Jimenez-Montano *et al.* (2002)
5–6	Telomeric/subtelomeric repeats	Kim and Wu (1997)
10–11	DNA bendability (helical repeat structure)	Fukushima *et al.* (2001) Herzel *et al.* (1999)
48–50	Centromeric repeats	Guy *et al.* (2003)
68	β satellite DNA	Waye and Willard (1989)
102	Nucleosomal structure in eukaryotes	Holste *et al.* (2003)
105–106	Isochores (DNA regions with low G + C content)	Buldyrev *et al.* (1995)
$\sim$135	Dimeric *Alu* repeat structure	Holste *et al.* (2003)
$\sim$165	Homopolymeric A-rich sequences within *Alu* repeats	Holste *et al.* (2003)
171	α satellite DNA	Haaf and Willard (1997)
$\sim$300	*Alu* repeats	Holste *et al.* (2003)
$\sim$680	DNA bend sites	Wada-Kiyama and Kiyama (1996)

is likely the most sophisticated and efficient information code created by nature through the process of evolution. However, while means of encoding some of this information is understood (e.g., the genetic code regulating amino acid assembly, directing intron/exon splicing), still relatively little is known about other layers of information encrypted in DNA.

According to Abel and Trevors (2005), DNA sequence complexity falls into three categories: (1) random sequence complexity, (2) ordered sequence complexity, and (3) functional sequence complexity.

(1) *Random sequence complexity* (in fact, *sequence randomness*) can be defined and measured solely in terms of probabilistic combinatorics, e.g., in terms of Shannon entropy. It has four components. (a) The number of "symbols" in the "alphabet" that could potentially occupy each location of the sequence. (b) Equal probabilistic availability of each "symbol" to each location. (c) The number of locations in the sequence. (d) Independence of each symbol from prior symbols.

(2) *Ordered sequence complexity* is a linear string of elements linked in patterns. It differs from random sequence complexity either by different probabilistic availability of each "symbol" or by dependence (correlation) of symbols.

(3) *Functional sequence complexity* is a linear string of symbols, where each symbol is a representation of a decision (selection) of a particular operation and a whole sequence can be defined as program (algorithm) for describing a certain sequence of actions (events). DNA sequences are not merely complex sequences, but they are *algorithmically* complex sequences, which could be analyzed not only as data, but also as programs.

3. Complexity Metrics

The complexity metrics, which can be used to analyze DNA sequences, can be broadly categorized into two groups: information-theoretic metrics (Shannon entropy, Rényi entropy, Kolmogorov complexity) and frequency analysis metrics (time delayed mutual information and autocorrelation function).

3.1. *Shannon Entropy*

Treating a discrete information source (e.g., DNA sequence) as a Markov process, Shannon entropy H can be used to measure how much information is generated by such process. The information source generates a series of symbols x_i belonging to an alphabet with size N. If symbols x_i are generated according to a known probability distribution $p(x_i)$, the entropy function $H(X)$ of n-length sequence $X = (x_1, x_2, \ldots, x_n)$, $x_i \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ can be defined as

$$H(X) = -\sum_{i=1}^{n} p(x_i) \cdot \log_2 p(x_i). \tag{1}$$

Shannon entropy is a measure of the uncertainty associated with a random variable. If the sequence consists of a series of symbols, entropy reaches a maximum, when the sequence satisfies the statistical criterion that all possible subsequences should appear with roughly equal probability and independent of prior options. In other words, entropy reaches maximum when the sequence is random. High entropy is an indicator of high randomness of a sequence, i.e., of a "genetic noise" (non-coding DNA sequences). Low entropy of a genomic sequence indicates low sequence complexity, which could be functionally important (Wan and Wootton, 2000; Chuzhanova *et al.*, 2003). Low complexity regions often containing some sequence repeats (Tautz *et al.*, 1986).

If the probabilities of symbol types are not equal at any position in a string then the information conveyed by the string is less than a maximum. The relative difference between the theoretical maximum of entropy $H_{\max}$ and the information contained in a given string is called *redundancy*. Thus, redundancy is defined as

$$R = 1 - \frac{H}{H_{\max}}. \tag{2}$$

If redundancy is close to zero, the sequence is random.

3.2. *Rényi Entropy*

Rényi entropy is a generalization of Shannon entropy for quantifying the diversity, uncertainty or randomness of a system. The Rényi entropy of order α, where $\alpha > 0$, is defined as

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \left(\sum_{i=1}^{n} p_i^\alpha(x_i) \right), \tag{3}$$

where p_i are the probabilities of $x_i \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$.

Second order Rényi entropy refers to the case $\alpha = 2$, which is derived from Eq. (3) as follows:

$$H_2(X) = -\log \sum_{i=1}^{n} p_i^2(x_i). \tag{4}$$

The Rényi entropies are important in ecology and statistics as indices of diversity. They also lead to a spectrum of indices of fractal dimension.

3.3. *Kolmogorov Complexity*

The main idea of Kolmogorov complexity is to measure the 'complexity' (i.e., information content) of an object by the length of the smallest program that generates it (Li and Vitanyi, 1997). In general case, we have a domain object X and a description system ϕ that maps from a description w (i.e., a program) to this object. Kolmogorov complexity $K_\phi(X)$ of an object X in the description system ϕ, is the length of the shortest program in the description system ϕ capable of producing X on a universal computer such as a Turing machine:

$$K_\varphi(X) = \min_w \big\{\|w\|\colon \varphi_w = X\big\}. \tag{5}$$

Kolmogorov complexity $K_\phi(X)$ is the minimal quantity of information required to generate X by an algorithm, and is the ultimate lower bound of information content. Unfortunately, it cannot be computed in the general case and must be approximated. Usually, compression algorithms are used to give an upper bound to Kolmogorov complexity. Suppose that we have a compression algorithm C_i. Then, a shortest compression of w in the description system ϕ will give the upper bound to information content in X:

$$K_\varphi(X) \leqslant C(X) := \min_i \big\{\|C_i\|, \varphi_{C_i} = X\big\}. \tag{6}$$

The relative Kolmogorov complexity (or compressibility) of sequence X can be calculated as follows:

$$\bar{C}(X) = \frac{C(X)}{\|X\|}, \tag{7}$$

where $\|X\|$ is the length of sequence X.

A sequence is considered random if it has no patterns and is incompressible ($\bar{C}(X) \approx 1$), i.e., we cannot find a shorter algorithm for specifying it than the sequence itself. If $\bar{C}(X) < 1$, this indicates the existence of some repeated patterns in a sequence.

3.4. *Autocorrelation*

The autocorrelation function (ACF) of a process describes the correlation between the parts of the process at different points in time (or position). Let X_t be the value of the

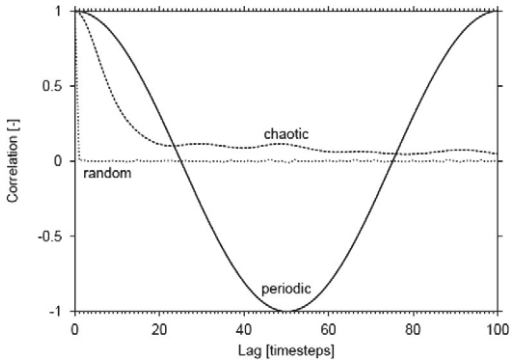

Fig. 1. Behavior of ACF for the periodic, chaotic, and random sequences (Daw *et al.*, 2003).

process at position t. If X_t has mean μ and variance σ^2 then the theoretical ACF is defined as follows:

$$R(t,s) = \frac{E[(X_t - \mu)(X_s - \mu)]}{\sigma^2}, \tag{8}$$

where E is the expected value operator.

If the statistical characteristics of X_t remain constant over time then the ACF depends only on the difference between t and s and can be expressed as a function $R(k)$ of a single variable k. This function $R(k)$ is an estimate of the theoretical ACF defined in Eq. (8) and is calculated as follows:

$$R(k) = \frac{1}{(n-k)\sigma^2} \sum_{t=1}^{n-k} [X_t - \mu]\,[X_{t+k} - \mu] \tag{9}$$

for any positive integer $k < n$, where k is the lag, and n is the length of the data series.

To calculate the ACF, the genomic data must be digitized first, i.e., the symbolic alphabet of 4 letters $\{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ must be converted to binary alphabet of 2 symbols $\{0, 1\}$, and whole sequence is treated as time series.

The typical behavior of the ACF for periodic, chaotic, and random sequences is presented in Fig. 1. For the random sequence, the ACF has a sharp peak at lag $= 0$, and the ACF value is almost zero for other *lag* values, indicating lack of correlation. For the chaotic sequence, the value of the ACF decays to a zero value after a finite lag. For the periodic sequence, the ACF has peaks, which correspond to certain periodicities in the data series (Daw *et al.*, 2003). Here a chaotic sequence is understood as a non-random sequence that appears to be random due to high sensitivity to nucleotide perturbations.

3.5. *Time Delayed Mutual Information*

The time delayed mutual information (TDMI) function (Fraser and Swinney, 1986) unlike the autocorrelation function also takes into account the non-linear correlations in the data

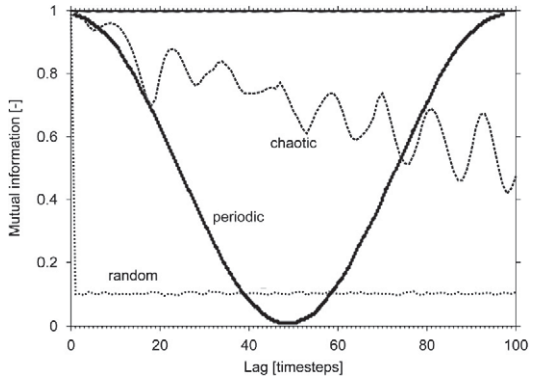

Fig. 2. Behavior of TDMI for the periodic, chaotic, and random sequences (Daw *et al.*, 2003).

series:

$$S = -\sum_{ij} p_{ij}(\tau) \ln \frac{p_{ij}(\tau)}{p_i p_j}, \tag{10}$$

where p_i is the probability to find a series value x in the i-th interval $[x_i, x_{i+1}]$, and $p_{ij}(\tau)$ is the joint probability that an observation falls into the i-th interval and the observation later, after delay (lag)τ, falls into the j-th interval.

The typical behavior of the TDMI function for periodic, chaotic, and random sequences is presented in Fig. 2. For the random sequence, the TDMI function has a sharp peak at *lag* $= 0$, and the TDMI value is almost zero for other *lag* values. For the chaotic sequence, the value of TDMI slowly decays to a zero value after a finite time. For the periodic sequence, the TDMI function shows peaks, which correspond to certain natural periodicities in the data series (Daw *et al.*, 2003).

4. Case Study: Analysis of the Human 22nd Chromosome

In a case study, we analyze DNA sequence of the human 22nd chromosome. The main properties of the 22nd chromosome are given in Table 2 (acc. to Holste *et al.*, 2001).

For analysis of the DNA sequence using Shannon entropy, the DNA sequence was split into 1000, 2000, 5000 and 10,000 bp windows, and Shannon entropy was calculated for each window using Eq. (1). The results were plotted against DNA sequence bp (see Fig. 3). Note that here for representation purposes DNA sequence bp is given for contiguous regions (DNA segments derived from a single genetic source) only (sequence gaps are excluded).

The graphs show that most of the DNA sequence fragments have entropy values close to 2 bits, which is maximum entropy for a message consisting of 4-letter alphabet ($\log_2(4) = 2$). It means that the distribution of symbols within such fragments is close to random and these fragments can be attributed to the non-coding regions of the DNA

Table 2
Main properties of human 22nd chromosome

Feature	Value
Length	33.8×10^6 (contiguous regions only)
G + C content	48%
Genes	546
Repeats	42%
No. of *Alu* repeats	22659
Alu repeats	17%

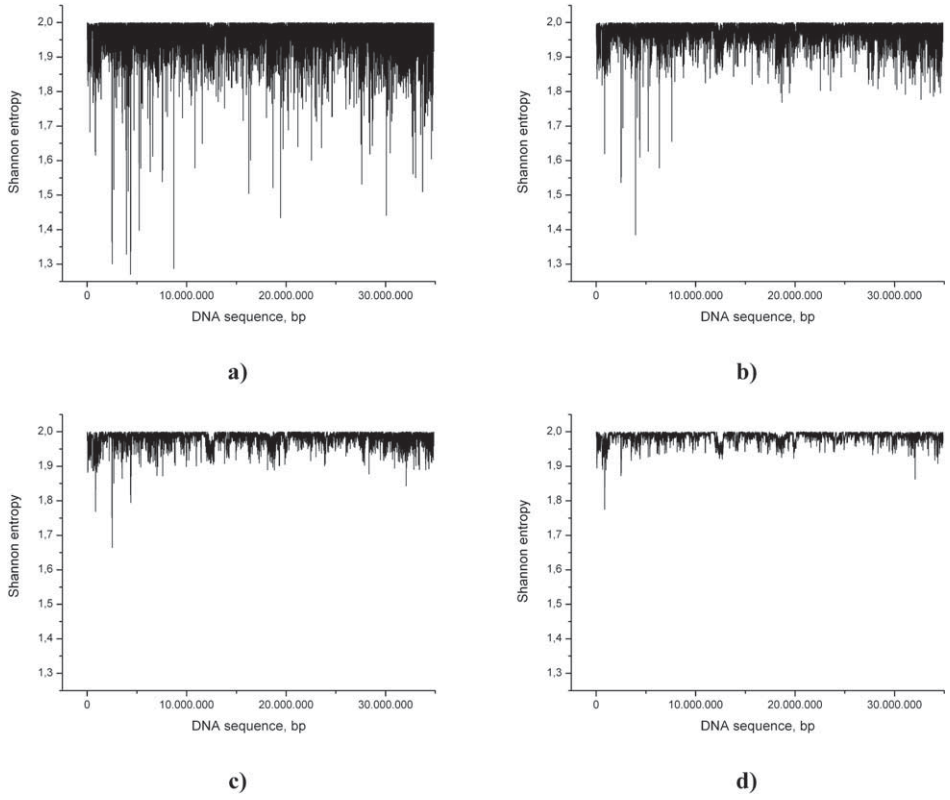

Fig. 3. Shannon entropy of DNA sequence: window size is (a) 1000 bp, (b) 2000 bp, (c) 5000 bp, and (d) 10,000 bp.

sequence ("genetic noise" or "junk DNA"). The regions, where entropy is significantly lower than 2 bits, point to lower complexity of the sequence. This lower entropy might be explained by unequal probabilities (non-uniform distribution) of nucleotides caused by repeats or patterns. Larger window size allows to smooth entropy graph and to isolate larger fragments with anomalous distribution of nucleotides.

For the entropy analysis of the DNA sequence using Rényi entropy, the DNA sequence was also split into 1000, 2000, 5000 and 10,000 bp windows, and the second-order Rényi entropy was calculated for each window using Eq. (4). The results were plotted against DNA sequence bp (see Fig. 4). The graphs show that similarly to Shannon entropy, most of the DNA sequence fragments have entropy values close to 2 bits. However, Rényi entropy is much more sensitive to lower complexity regions, which allows identifying them more easily.

For the estimation of relative Kolmogorov complexity, we use a *zzip* compressor that implements a version of Burrows-Wheeler Transform (BWT) algorithm, which is a block-sorting algorithm particularly suitable for compressing text information. The DNA sequence was split into 1000, 2000, 5000 and 10,000 bp windows, which were compressed, the compressibility of the sequence was calculated using Eq. (7) and the results presented in Fig. 5. The theoretical maximum of the metric is 0.25, because 8-bit text symbols can be compressed to 2 bits, i.e., we can achieve a compression factor of 4. In practice, the compressibility values can be higher than 0.25 due to the compression overhead. For regions with low complexity and repeating patterns of nucleotides, relative Kolmogorov complexity can be as much as 2 times lower than a maximal value (see Fig. 5a). Furthermore, the established regions with low complexity are not the same regions as were obtained using Shannon and Rényi entropies.

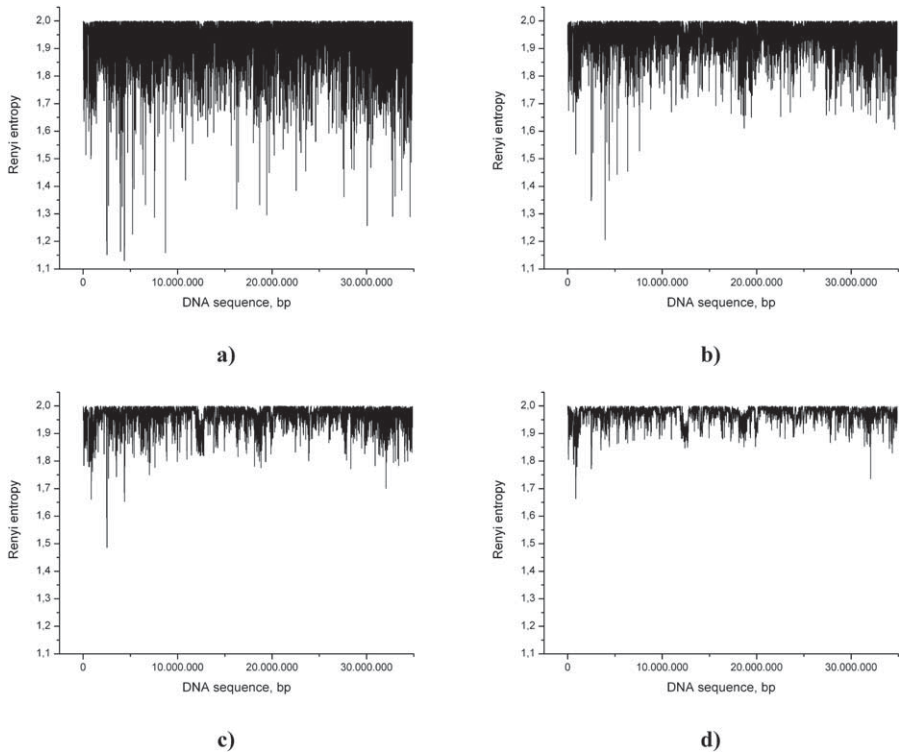

Fig. 4. Rényi entropy of DNA sequence: window size is (a) 1000 bp, (b) 2000 bp, (c) 5000 bp, (d) 10,000 bp.

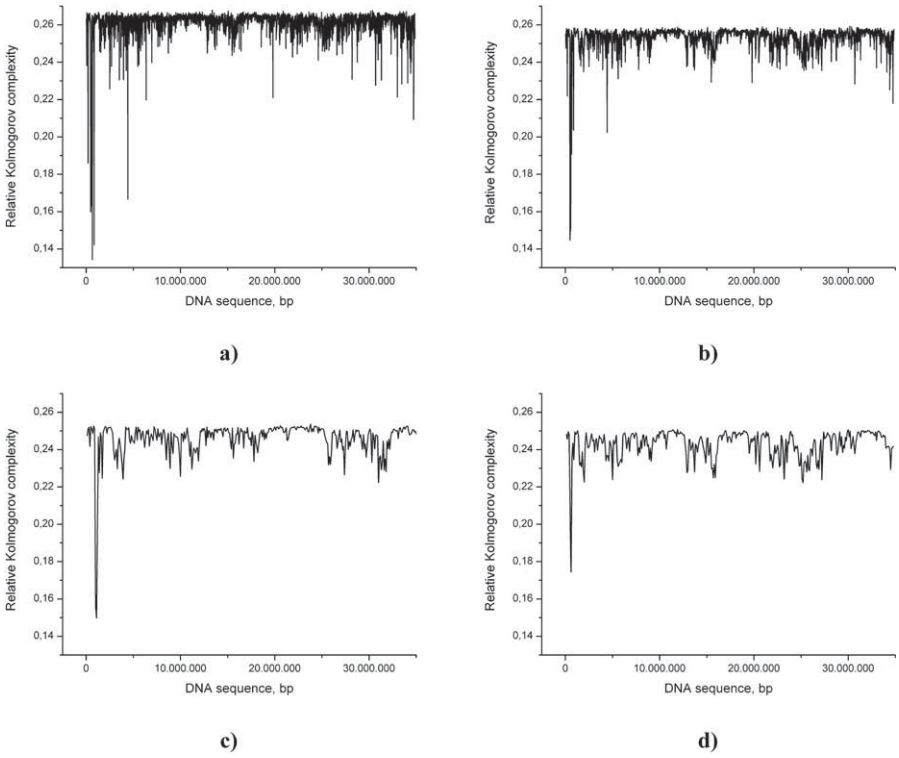

Fig. 5. Relative Kolmogorov complexity of DNA sequence: window size (a) 1000 bp, (b) 2000 bp, (c) 5000 bp, (d) 10,000 bp.

The comparison of results obtained using the information-theoretic metrics is given in Table 3.

Next, we perform the frequency analysis of the DNA sequence. The sequence was converted into binary code using simple mapping rules $\{\mathrm{A}, \mathrm{T}\} \rightarrow 1$ and$\{\mathrm{G}, \mathrm{C}\} \rightarrow 0$.

Table 3
Summary of DNA analysis using information-theoretic metrics

Metric	Min. value in sequence fragment, when window size is			
	1000 bp	2000 bp	5000 bp	10,000 bp
Shannon entropy	(4,358,000–4,359,000) **1.2706**	(3,950,000–3,952,000) **1.3850**	(2,515,000–2,520,000) **1.6647**	(840,000–850,000) **1.7755**
Rényi entropy	(2,520,000–2,521,000) **1.1517**	(3,950,000–3,952,000) **1.2061**	(2,515,000–2,520,000) **1.4866**	(840,000–850,000) **1.6641**
Relative Kolmogorov complexity	(659,000–660,000) **0.1343**	(478,000–500,000) **0.1447**	(545,000–550,000) **0.1497**	(590,000–600,000) **1.1744**

Then, Eq. (9) and Eq. (10) were used to calculate the ACF and TDMI values, which were plotted against different lags (see Fig. 6). The results show that the entire DNA sequence is not random, however the data series has a chaotic behavior with exponentially decaying ACF and TDMI values. Only a periodicity of $f^{-1} = 3$ bp could be identified for sure.

Therefore, the straightforward application of the frequency analysis methods is not beneficial, because the entire DNA sequence consists of many regions with different periodicities, which suppress each other and as a result, only the chaotic behavior of the sequence could be established. For deeper analysis, we must select smaller DNA regions.

Fur further analysis, based on Table 3 we select two interesting regions: (1) 3,900,000–4,000,000 bp (for simplicity, we denote it as Region 1); and (2) 600,000–700,000 bp (Region 2). The size of both regions is 100 Kbp. Region 1 was established using both Shannon and Rényi entropy, while Region 2 was established using relative Kolmogorov complexity. The results of the frequency analysis of Regions 1 and 2 are presented in Figs. 7 and 8, respectively.

In Fig. 7, a periodicity of $f^{-1} = 3$ bp is clearly seen, as both the autocorrelation function and time delayed mutual information (TDMI) have maximal values at multiples of 3 bp (3, 6, 9 bp, etc.) It means that the analyzed fragment likely contains the protein coding DNA sequence, i.e., the gene(s).

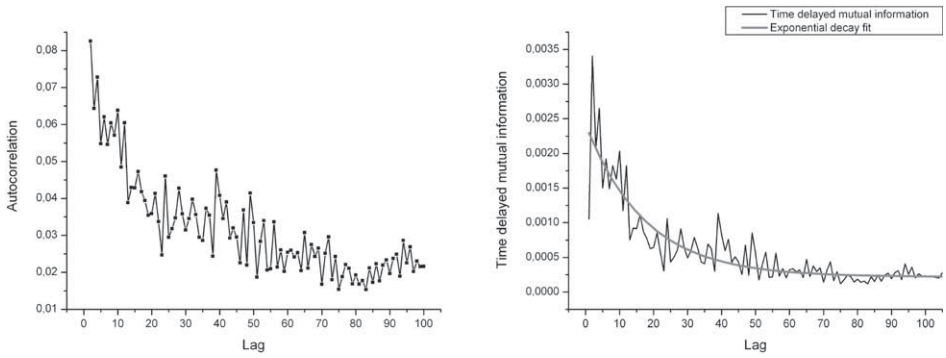

Fig. 6. Autocorrelation (left) and time delayed mutual information (right) of the entire DNA sequence.

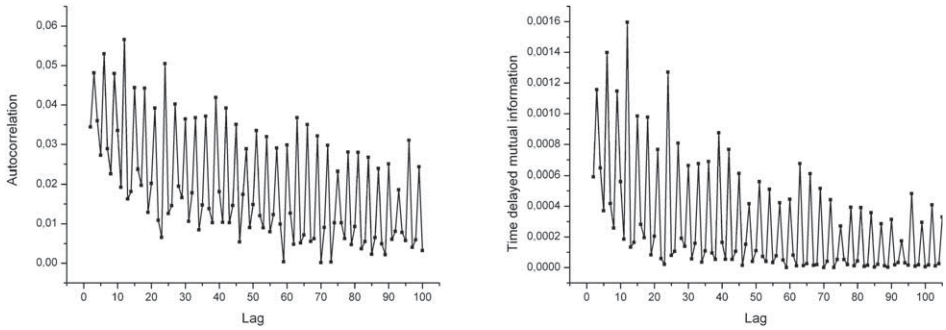

Fig. 7. Analysis of DNA Region 1: autocorrelation (left) and TDMI (right).

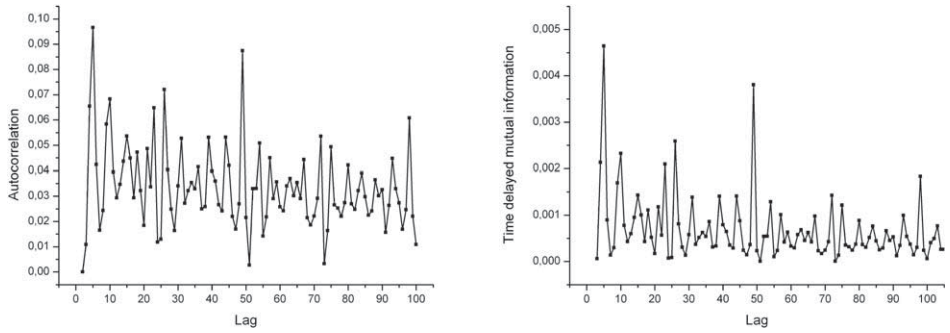

Fig. 8. Analysis of DNA Region 2: autocorrelation (left) and TDMI (right).

In Fig. 8, the periodicities of $f^{-1} = 5$ bp, which is characteristic to subtelomere repeats, and $f^{-1} = 49$ bp (and its multiples: 98 bp, 147 bp, 196 bp, ...) are clearly seen, which means that this DNA sequence fragment can be attributed to a peri-centromeric region. This region has no noticeable $f^{-1} = 3$ bp periodicity. Therefore, we can assume that this sequence fragment contains no or very few protein coding genes.

5. Evaluation of Results

We can compare our analysis results with the actual database annotations of the 22nd chromosome sequence given in the NCBI database (NCBI).

Region 1 (actual sequence position with gaps included: 18,480,000–18,580,000 bp) contains 3 genes (RANBP1, ZDHHC8 and LOC388849), which in total have 37,289 bp length, i.e., they occupy more than 37% of Region 1. That also confirms our analysis results that this DNA sequence fragment is rich in protein-coding genes.

Region 2 (actual sequence position with gaps included: 15,030,000–15,280,000 bp) contains only 1 gene (ABCD1P4), which is only 782 bp length, i.e., it occupies less than 0.8% of Region 2. That confirms our analysis results that this DNA sequence fragment is poor in protein-coding genes.

Furthermore, Region 2 is close to the centromere of the chromosome as predicted by our analysis results. The centromere is the thinnest region of a chromosome that plays a role in cellular division and the control of gene expression. The centromere has no defined DNA sequence, and it typically consists of large arrays of repetitive DNA (e.g., satellite DNA), which can be detected using frequency analysis of DNA sequence.

Region 2 also belongs to the subtelomere of the 22nd chromosome. The subtelomeric region is the chromosomal region just proximal to the telomere (the end of the chromosome) and composed of highly polymorphic repetitive DNA sequences that are typically situated adjacent to the gene-rich areas. Identification of the subtelomeric regions is very important, because mutations in these regions can cause mental retardation.

6. Conclusions

Genetic sequences have very complex structure with different regions, which have different periodicities or exhibit random behavior. The analysis of such sequences cannot be performed with satisfactory results using a single method. Therefore, a combination of different methods should be used. This paper proposes to analyze the DNA sequences in two stages.

First, the information-theoretic methods (Shannon entropy, Rényi entropy, relative Kolmogorov complexity) can be used to estimate their randomness or, inversely, their repeatability and to detect smaller regions of particular interest (with lower complexity). The results of a case study show that both coding and non-coding sequences can have low complexity; and relative Kolmogorov complexity and entropies can establish different low complexity regions. However, the information-theoretic methods do not measure the structural organization of the sequence.

Second, the identified regions with low complexity can be further analyzed using the frequency analysis methods such as autocorrelation or time delayed mutual information. The latter methods allow detecting specific periodicities and other important structural features of the DNA sequences such as telomers, subtelomers or gene-rich fragments.

References

Abel, D.L., Trevors, J.T. (2005). Three subsets of sequence complexity and their relevance to biopolymeric information. *Theor. Biol. Med. Model.*, 2, 29.

Adjeroh, D., Zhang, Y., Mukherjee, A., Powell, M., Bell, T. (2002). DNA sequence compression using the burrows-wheeler transform. In: *Proc. IEEE Comput. Soc. Bioinform. Conf.*, 1, pp. 303–313.

Azbel, M.Y. (1995). Universality in a DNA statistical structure. *Phys. Rev. Lett.*, 75(1), 168–171.

Berger, J.A., Mitra, S.K. (2002). New approaches to genome sequence analysis based on digital signal processing. In: *Workshop on Genomic Signal Processing and Statistics* (*GENSIPS–2002*). Raleigh, North Carolina, USA.

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.

Chang, C.H., Hsieh, L.C., Chen, T.Y., Chen, H.D., Luo, L.F., Lee, H.C. (2005). Shannon information in complete genomes. *J. Bioinform. Comput. Biol.*, 3, 587–608.

Chen, X., Kwong, S., Li, M.A. (1999). Compression algorithm for DNA sequences and its applications in genome comparison. In: *Genome Inform. Ser. Workshop Genome Inform.*, 10, 51–61.

Chen, H.D., Chang, C.H., Hsieh, L.C., Lee, H.C. (2005). Divergence and Shannon information in genomes. *Phys. Rev. Lett.*, 94(17), 178103.

Chuzhanova, N.A., Anassis, E.J., Ball, E., Krawczak, M., Cooper, D.N. (2003). Meta-analysis of indels causing human genetic disease: mechanisms of mutagenesis and the role of local DNA sequence complexity. *Hum. Mutat.*, 21, 28–44.

Crisan, M. (2004). On Shannon, Fisher, and algorithmic entropy in cognitive systems. In: *1st Romanian-Hungarian Joint Symposium on Applied Computational Intelligence SACI-2004*. Timisoara, Romania, pp. 49–56.

Cristea, P.D., Popescu, G.A. (2003). Fractal dimension of human chromosome 22. In: *Proc. of 1st South–East European Symposium on Interdisciplinary Approaches in Fractal Analysis* (*IAFA '03*). Bucharest, Romania, pp. 131–134.

Daw, C.S., Finney, C.E.A., Tracy, E.R. (2003). A review of symbolic analysis of experimental data. *Rev. Sci. Instrum.*, 74(2), 915–930.

De Sousa Vieira, M. (1999). Statistics of DNA sequences: A low-frequency analysis. *Phys. Rev. E*, 60, 5932–5937.

Dunham, I. *et al.* (1999). The DNA sequence of human chromosome 22, *Nature*, 402, 489–495.

Frappat, L., Minichini, C., Sciarrino, A., Sorba, P. (2003). Universality and Shannon entropy of codon usage. *Phys. Rev. E. Stat. Nonlin. Soft. Matter Phys.*, 68(6 Pt 1), 061910.

Fraser, A.M., Swinney, H.L. (1986). Independent coordinates for strange attractors from mutual information. *Phys. Rev. A*, 33, 1134–1140.

Freeman, J., Plasterer, T.N., Smith, T.F., Mohr, S.C. (1998). Patterns of genome organization in bacteria. *Sci. Mag.*, 279(5358), 1827.

Fukushima, A., Kinouchi, M., Kudo, Y., Kanaya, S., Mori, H., Ikemura, T. (2001). Statistical analysis of genomic information: Various periodicities in DNA Sequence. *Genome Inform.*, 12, 435–436.

Fukushima, A., Ikemura, T., Oshima, T., Mori, H. (2002). Detection of period in eukaryotic genomes on the basis of power spectrum analysis. *Genome Inform. Ser. Workshop Genome Inform.*, 13, 21–29.

Grosse I., Herzel, H., Buldyrev, S.V., Stanley, H.E. (2000). Species independence of mutual information in coding and non-coding DNA. *Phys. Rev. E*, 61, 5624–5629.

Gusev, V.D., Kulichkov, V.A., Chupakhina, O.M. (1993). The Lempel–Ziv complexity and local structure analysis of genomes. *Biosystems*, 30(1–3), 183–200.

Guy, J., Hearn, T., Crosier, M., Mudge, J., Viggiano, L., Koczan, D., Thiesen, H.–J., Bailey, J.A., Horvath, J.E., Eichler, E.E., Earthrow, M.E., Deloukas, P., French, L., Rogers, J., Bentley, D., Jackson, M.S. (2003). Genomic sequence and transcriptional profile of the boundary between pericentromeric satellites and genes on human chromosome arm 10 p. *Genome Res.*, 13(2), 159–172.

Haaf, T., Willard, H.F. (1997). Chromosome-specific a-satellite DNA from the centromere of chimpanzee chromosome 4. *Chromosoma*, 106(4), 226–232.

Hancock, J.M., Armstrong, J.S. (1994). SIMPLE34: an improved and enhanced implementation for VAX and SUN computers of the SIMPLE algorithm for analysis of clustered repetitive motifs in nucleotide sequences. *CABIOS*, 10, 67–70.

Herzel, H., Ebeling, W., Schmitt, A.O. (1994). Entropies of biosequences: the role of repeats. *Phys. Rev. E*, 50, 5061–5071.

Herzel, H., Trifonov, E.N., Weiss, O., Grobe, I. (1998). Interpreting correlations in biosequences. *Physica A*, 248, 449–459.

Herzel, H., Weiss, O., Trifonov, E.N. (1999). 10–11 bp periodicities in complete genomes reflect protein structure and DNA folding. *Bioinformatics*, 15, 187–193.

Holste D., Grosse, I., Herzel, H. (2001). Statistical analysis of the DNA sequence of human chromosome 22. *Phys. Rev. E*, 64, 1–9.

Holste, D., Grosse, I., Beirer, S., Schieg, P., Herzel, H. (2003). Repeats and correlations in human DNA sequences. *Phys. Rev. E*, 67, 061913.

Hou, J. (2007). A matrix-based model for web page community construction and more. *Informatica*, 18(2), 217–238.

Jimenez-Montano, M.A., Ebeling, W., Pohl, T., Rapp, P.E. (2002). Entropy and complexity of finite sequences as fluctuating quantities. *Biosystems*, 64, 23–32.

Kim, N.W., Wu, F. (1997). Advances in quantification and characterization of telomerase activity by the telomeric repeat amplification protocol (TRAP). *Nucleic Acids Res.*, 25(13), 2595–2597.

Larsabal, E., Danchin, A. (2005). Genomes are covered with ubiquitous 11 bp periodic patterns, the "class A" flexible patterns. *BMC Bioinform.*, 6, 206.

Li, W. (1997). The study of correlation structures of DNA sequences: A critical review. *J. Comput. Chem.*, 21, 257–271.

Li, W., Kaneko, K. (1992). Long-range correlation and partial 1/f spectrum in a noncoding DNA sequence. *Europhys. Lett.*, 17, 655–660.

Li, M., Vitanyi, P. (1997). *An Introduction to Kolmogorov Complexity and Its Applications*. Springer-Verlag.

Lyshevski, S.E., Krueger, F.A. (2004). Nanoengineering bioinformatics: Fourier transform and entropy analysis. In: *Proc. of the 2004 American Control Conference*. Boston, Massachusetts, pp. 317–322.

Mansilla, R., Del Castillo, N., Govezensky, T., Miramontes, P., Jose, M., Coch, G. (2004). Long-range correlation in the whole human genome. In: *Workshop on Interdisciplinary Approaches to Some Complex Biological Systems*. Mexico City, Mexico.

Mantegna, R.N., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Peng, C.K., Simons, M., Stanley, H.E. (1994). Linguistic features of non-coding DNA sequences. *Phys. Rev. Lett.*, 73, 3169–3172.

NCBI. Online: `http://www.ncbi.nlm.nih.gov/mapview/maps.cgi?TAXID=9606&CHR=22`

Nowak, R. (1994). Mining treasures from 'Junk DNA'. *Science*, 263(5147), 608–610.

Orlov, Y.L., Filippov, V.P., Potapov, V.N., Kolchanov, N.A. (2002). Construction of stochastic context trees for genetic texts. *In Silico Biol.*, 2, 233–247.

Peng, C.K., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Mantegna, R.N., Simons, M., Stanley, H.E. (1995). Statistical properties of DNA sequences. *Physica A*, 221, 180–92.

Schmitt, A.O., Herzel, H. (1997). Estimating the entropy of DNA sequences. *J. Theor. Biol.*, 188(3), 369–377.

Schurmann, T., Grassberger, P. (1996). Entropy estimation of symbol sequences. *Chaos*, 6(3), 414–427.

Stanley, H.E., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Peng, C.K., Simons, M. (1999). Scaling features of noncoding DNA. *Physica A*, 273, 1–18.

Tautz, D., Trick, M., Dover, G.A. (1986). Cryptic simplicity in DNA is a major source of genetic variation. *Nature*, 322, 652–656.

Troyanskaya, O.G., Arbell, O., Koren, Y., Landau, G.M., Bolshoy, A. (2002). Sequence complexity profiles of prokaryotic genomic sequences: A fast algorithm for calculating linguistic complexity. *Bioinformatics*, 18, 679–688.

Vinga, S., Almeida, J.S. (2004). Rényi continuous entropy of DNA sequences. *J. Theor. Biol.*, 231(3), 377–388.

Voss, R.F. (1992). Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. *Phys. Rev. Lett.*, 68, 3805–3808.

Wada-Kiyama, Y., Kiyama, R. (1996). Conservation and periodicity of DNA bend sites in eukaryotic genomes. *DNA Res.*, 3, 25–30.

Wan, H., Wootton, J.C. (2000). A global compositional complexity measure for biological sequences: AT-rich and GC-rich genomes encode less complex proteins. *Comput. Chem.*, 24, 71–94.

Waye, J.S., Willard, H.F. (1989). Human beta satellite DNA: genomic organization and sequence definition of a class of highly repetitive tandem DNA. *Proc. Natl. Acad. Sci. USA*, 86(16), 6250–6254.

Wootton, J.C., Federhen, S. (1996). Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.*, 266, 554–71.

Yu, Z.-G., Anh, V. (2001). Time series model based on global structure of complete genome. *Chaos Soliton Fract.*, 12(10), 1827–1834.

R. Damaševičius received the PhD degree in 2005 in informatics from Kaunas University of Technology, Kaunas, Lithuania. Currently he is an assoc. prof. at Software Engineering Department, Kaunas University of Technology. His research interests include software reuse, software generation and program transformation, as well as bioinformatics and data mining.

Genetinių sekų sudėtingumo įvertinimas naudojant informacijos teorijos ir dažnių analizės metodus

Robertas DAMAŠEVIČIUS

Genetinė informacija ląstelėse saugoma DNR sekoje, kuri yra užrašoma kaip iš keturių skirtingų nukleotidų tipų simbolių sudaryta eilutė (seka). Genetinėse sekose gausu pasikartojimų, kurie yra biologiškai svarbūs. Genetinių sekų sudėtingumą galima įvertinti naudojant informacijos teorijos metodus. Žemo DNR sekos regiono sudėtingumas rodo, kad jame yra pasikartojančių sekos fragmentų. Šiame straipsnyje genetinių sekų sudėtingumas yra įvertinamas naudojant Šenono entropiją, Renyi entropiją ir santykinį Kolmogorovo sudėtingumą. Struktūrinis sekų sudėtingumas yra analizuojamas naudojant autokoreliacijos funkciją ir uždelstą bendrąją informaciją. Šios metrikos taikomos žmogaus 22-osios chromosomos genetinės sekos tyrimui.