REVIEWS OF TOPICAL PROBLEMS PACS numbers: **02.50.–r**, **05.10.–a**, **87.10.+e**, 87.14.Gg

Order and correlations in genomic DNA sequences. The spectral approach

V V Lobzin, V R Chechetkin

DOI: 10.1070/PU2000v043n01ABEH000611

Contents

Abstract. **The structural analysis of genomic DNA sequences is discussed in the framework of the spectral approach, which is sufficiently universal due to the reciprocal correspondence and mutual complementarity of Fourier transform length scales. The spectral characteristics of random sequences of the same nucleotide composition possess the property of self-averaging for relatively short sequences of length $M \geqslant 100-300$. Comparison with the characteristics of random sequences determines the statistical significance of the structural features observed. Apart from traditional applications to the search for hidden periodicities, spectral methods are also efficient in studying mutual correlations in DNA sequences. By combining spectra for structure factors and correlation functions, not only integral correlations can be estimated but also their origin identified. Using the structural spectral entropy approach, the regularity of a sequence can be quantitatively assessed. A brief introduction to the problem is also presented and other major methods of DNA sequence analysis described.**

V V Lobzin Institute of Terrestrial Magnetism, Ionosphere,
and Radiowave Propagation, Russian Academy of Sciences
142092 Troitsk, Moscow Region, Russian Federation
Tel. (7-095) 334 01 13. Fax (7-095) 334 01 24
E-mail: lobzin@top.izmiran.troitsk.ru
V R Chechetkin Troitsk Institute for Innovation and Thermonuclear Investigations, 142092 Troitsk, Moscow Region, Russian Federation
Tel. (7-095) 334 50 57 E-mail: vladimir_che@mail.ru

Received 1 June, revised 30 July 1999
Uspekhi Fizicheskikh Nauk **170** (1) 57 – 81 (2000)
Translated by V V Lobzin; edited by M S Aksent'eva

1. Introduction

The rapid growth in the number of determined genomic DNA sequences and the increase in the speed and memory capacity of modern computers are two of the most typical features of science at the end of the 20th century. By March 1999 the databases contained 3.3×10^6 sequences comprising 2.4×10^9 nucleotides. The tendency of doubling the data each 1 – 1.5 years is expected to persist in the near future (see Fig. 1). By 2003 the complete human genomic DNA containing $\sim 3 \times 10^9$ nucleotides should be determined. Processing this information requires the combined efforts of not only geneticists, biologists, biochemists, and physicians but mathematicians and programmers as well. From the early 90s, physicists have also been involved in this activity, and papers on the subject began to appear in the leading physical journals. It is worth noting that the laboratories in Los-Alamos and Livermore, for example, are leaders not only in physical research but also in the computer analysis of DNA sequences.

The information stored in genomic DNA should be reproduced, recognized, read, and serve as a specific program for triggering a number of molecular mechanisms. Thus the functions of various sites in a DNA sequence may also be essentially different. A primary task in the computer analysis of DNA sequences is to reveal the reproducible structural features and to relate them to the corresponding functions. A variety of ways of molecular evolution and the distinctions in the environment conditions for different organisms result in a number of functional-structural relationships. Selection and fixing such relationships during molecular evolution allows the use of a system approach based on the comparative analysis of DNA sequences in the databases.

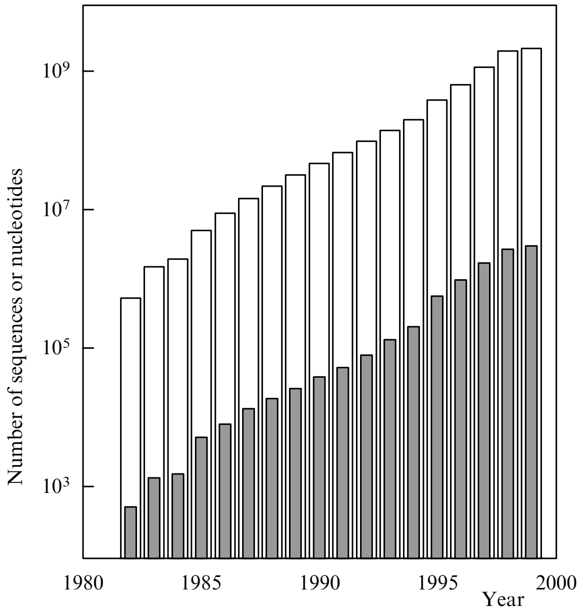

Figure 1. Increase in the number of DNA sequences determined (shaded bars) and nucleotides in them (white bars).

To begin with, we describe the databases because almost all the data (apart from some results ordered by pharmaceutical corporations) are freely accessible via the Internet. A comprehensive set of data on genomic DNA sequences is contained in GenBank (Los-Alamos, USA) and the databases of the European molecular biological laboratory (EMBL, Heidelberg, Germany) and European Bioinformatics Institute (a EMBL outstation located near Cambridge, England). In the general purpose databanks, an accession number is assigned to each DNA sequence. Apart from the general purpose databanks, there are more than fifty specialized databases. A detailed description of them and their URLs (uniform resource locators) can be found in Refs [1 – 3].

The bibliography of papers in Russian, which can be used as an introduction to the subject concerned, is not very extensive [4 – 7] and corresponds to the state-of-art up to 1990. From the plenty of relatively recent publications we could recommend [8 – 15]. The current bibliography of papers published since 1992 and oriented to the physical audience can be found at URL [16].

The subject of the review is primarily determined by our own studies [17 – 27] devoted to spectral methods for structural analysis of DNA sequences. Nevertheless, in the introductive Section 2 we tried at least briefly to outline the other main methods for analysis of DNA sequences. The general scheme for spectral analysis is given in Section 3. As the applications of the general theory, the search for hidden periodicities and analysis of correlations in DNA sequences are considered in Sections 4 and 5, respectively. The variety of ways of molecular evolution results in significant structural distinctions even in the functionally similar DNA segments for different species. As an illustration, we selected only a few examples but tried to show the reproducible and to some extent universal structural features.

Before going to the main content of the review, we try to answer the following questions: Why is it the spectral approach that we use? What are the advantages and drawbacks of the approach? Since the Fourier transform is reversible, from the formal viewpoint we can talk about the one-to-one mapping of information with the use of Fourier transform. The mutual complementarity of scales for Fourier transform allows one to overlap all the characteristic lengths. This approach also allows unification of the description of the characteristics not only for symbolic sequences but for spatial linear chains as well (see Conclusions). Thus using the spectral approach, we obtain an opportunity to advance a little farther in studying the relationship between DNA sequences and protein structures.

As a rule, the studies of structural relationships are closely connected with the problems of molecular evolution. In Ref. [28] a scenario is considered involving a block-hierarchical assembly and selection of structural units during molecular evolution. The spectral analysis allows the development of such a scenario in more detail.

Frequently, the statistical significance of the regular relationships observed turns out rather low, and the verification of the relationships often requires a combination of different methods (see Section 5). Thus, although the spectral approach is quite universal, in practice it is impossible to use this method alone. We need the much larger arsenal of methods partially described in Section 2.

2. Genomic DNA sequences and analysis of them

From the physical viewpoint, the study of mutual correlations, periodicities, and commensurability – incommensurability effects in DNA sequences are of the greatest interest. To interpret these effects, one should be acquainted with the structure and molecular biological processes DNA is involved in.

2.1 Brief molecular biology review

The genetic information on the development of an organism is stored in the long biopolymeric nucleic acid macromolecules [29 – 31]. The structural subunits of the nucleic acids are called nucleotides, which are large molecular complexes consisting of a nitrogen-containing base, a five-carbon sugar residuum, and a phosphate group. In accordance with the two kinds of sugars, there exist desoxyribonucleic acid (DNA) and ribonucleic acid (RNA). A DNA molecule contains bases of four kinds, namely, adenine (A), guanine (G), cytosine (C), and thymine (T). In RNA, thymines are replaced by uracils (U). In pentose rings of sugars the carbons are numbered in accordance with a standard rule. During polymerization, a covalent bond is formed between the $3'$ group of the sugar residuum of one nucleotide and the $5'$ group of another. Therefore the beginning of the nucleotide sequence is called the $5'$ end, and its termination the $3'$ end. Apart from some viruses, the genetic information is stored in double-stranded DNA in which each adenine is bound with a thymine by two hydrogen bonds and each guanine is bound with a cytosine by three hydrogen bonds.

During analysis, different nucleotides are often combined into binary groups, for example, purines $\mathrm{R} = (\mathrm{A}, \mathrm{G})$ and pyrimidines $\mathrm{Y} = (\mathrm{C}, \mathrm{T}$ or $\mathrm{U})$, as well as the groups differing in the bond strength, $\mathrm{W} = (\mathrm{A}, \mathrm{T})$ and $\mathrm{S} = (\mathrm{C}, \mathrm{G})$, or in the physico-chemical properties, the so-called ketone-amino subdivision, $\mathrm{K} = (\mathrm{G}, \mathrm{T})$ and $\mathrm{M} = (\mathrm{A}, \mathrm{C})$.

Double-stranded DNA can exist in an A form, which is a right-handed helix with a pitch of $l \simeq 10.8$ base pairs (in the following l is measured in base pairs for double-stranded molecules and in bases for single-stranded molecules), in a B form, which is also a right-handed helix ($l \simeq 10.2-10.5$), and a Z form, which is a left-handed helix ($l \simeq 12.0$). The characteristic feature of the Z form is the alternation of purines and pyrimidines, RYRY... Commonly, DNA is in the B form, however, in genomic DNA the local dynamic transitions B–A and B–Z may occur under certain conditions (depending on the salt concentrations, the level of supercoiling, etc.).

Biologists make a distinction between prokaryotic and eukaryotic cells, the latter being at a higher level of evolution. Unlike the prokaryotic cells, the eukaryotic cells have a nucleus ('karyon' means 'nucleus' in ancient Greek). The compact DNA packing in a nucleus is achieved by a specific hierarchical folding. At the first level of compactization the eukaryotic DNA is assembled into nucleosome units where fragments of 146 base pairs are wound around protein molecular complexes composed of histones. Between the DNA fragments fixed in such a manner there are connective DNA fragments. The length of DNA in the nucleosomes varies within the range $l = 200 \pm 40$. Then, the nucleosomes are assembled into a 30 nm solenoid-like fibril. The pitch of the solenoid is of about six nucleosomes. On the third level of compactization the loops of fibrils are formed. The length of the loops varies over a wide range, $l \simeq 2 \times 10^4 - 10^5$. Finally, on the fourth level of compactization the loops are stowed. Within the quasi-regular stowage, units composed of ~ 10 loops can be distinguished.

The bending rigidity of double-stranded DNA is characterized in terms of Kuhn length, l_{Kuhn} [32]. For typical physiological conditions $l_{\text{Kuhn}} \simeq 300-340$. Instead, the sometimes persistent length is used, $l_{\text{pers}} = 0.5 l_{\text{Kuhn}}$. It is seen that for the mean length of a DNA fragment in a nucleosome, l_{nucl}, the inequalities $l_{\text{pers}} < l_{\text{nucl}} < l_{\text{Kuhn}}$ hold. The Kuhn length for single-stranded DNA is much smaller, $l_{\text{Kuhn}} \simeq 12-14$.

In prokaryotic organisms a considerable amount of DNA codes for proteins. First, using the coding DNA as a template, an RNA, which is called a messenger RNA or mRNA, is synthesized. This process is called a transcription. Further, the protein molecule is synthesized on mRNA in accordance with the universal genetic code. This process is called translation. During translation, each nucleotide triplet corresponds to one amino acid, and the sequence of amino acids forms a protein molecule. Usually only one of the DNA strands codes for proteins. The beginning of translation is determined by the initiating codone (as a rule, ATG) and a specifically recognized region preceding the initiating codone. The end of translation is determined by the terminating codones (TAA, TAG or TGA). Since there exist only 20 different amino acids, the genetic code is degenerate.

The three-dimensional structure of proteins also has a strongly pronounced hierarchical character (see, e.g., Ref. [33]). The primary level (or a primary structure) is usually identified with the amino acid sequence. The secondary structure of proteins characterizes the spatial arrangement of atoms within separate regions of the main chain of a protein molecule, and the tertiary structure is the characteristic manner of space folding for the entire protein molecule [33, 34]. The basic elements of secondary structure are α-helices and approximately flat β-sheets [32–34]. The pitch of an α-helix is composed of approximately 3.6 ± 0.2 amino acids, and the characteristic property of the β-sheets is the alternation of hydrogen bonds. As a rule, the elements of secondary structure are composed of $\sim 8-20$ amino acids and are connected with each other by connective coil-like fragments. The total number of amino acids in protein molecules varies within the range $\sim 70-2000$ ($\sim 200-300$ on the average).

In eukaryotic genomes there are three classes of DNA sequences [29, 31, 35], namely, a satellite DNA consisting of multiple repeats, moderately repetitive sequences scattered over the genome, and unique sequences. The satellite DNA serves for the correct structural organization of chromosomes. The moderately repetitive sequences play in part a structural role and in part a regulatory role. Some families of repeats are so specific that they can be used as 'genetic fingerprints' [6]. A fraction of the unique DNA codes for proteins. In mammals, the fraction of unique DNA is rather small, $\sim 2-3\%$. A characteristic feature of the eukaryotic protein-coding DNA is its split character, i.e., the coding regions (exons) alternate with the non-coding regions (introns). At the first stage, the corresponding RNA copy is wound around a protein core thereby forming ribonucleoprotein particles (RNP particles), which consist of RNA fragments of length ~ 600 nucleotides and are joined by connective fragments of length ~ 100 bases. Then introns are snipped out, and stitched exons form an mRNA. This process is called splicing. The proportions between the three above-mentioned classes of DNA sequences vary over a wide range, $\sim 10-50\%$, for different species.

Dynamically binding with DNA (the process is often cooperative), the regulatory proteins change the character of chromatin packing and 'turn on' (or 'turn off') different genes [29–31]. The lengths of contacts vary within the range $l_{\text{contact}} \sim 5-100$.

Some characteristic lengths for eukaryotic genomes are related to other dynamic processes such as a site-specific recombination, when DNA fragments are incorporated into and excised from the genome, and crossing over, when an exchange of fragments between homologous chromosomes occurs [29]. These lengths vary over a very wide range and require a separate analysis.

A complete list of the characteristic lengths is given in Table 1. Some characteristic lengths are quite conservative, while the others vary within rather wide ranges. It is essential that the peculiarities of a DNA molecular structure, the packing mechanisms, and the processes, in which DNA is involved, are reflected in the DNA sequences in one way or another [36, 37]. The differences in the structure and physicochemical properties of different bases result in the energetic and functional non-equivalence of the double-stranded DNA molecules with different nucleotide sequences. The sequences acquiring some energetic and/or functional advantages, have, in turn, a preference in the process of evolutionary selection. For example, regular insertions of fragments consisting of several adenines result in a bend of double-stranded DNA [38–40]. It is easier to pack such DNA into nucleosomes. On the other hand, one could expect a quasi-regular packing in nucleosomes be displayed as a hidden periodicity in the DNA sequences (this effect is really observed [26]).

The functional distinctions between different DNA segments result in a characteristic heterogeneity of the structure and structural characteristics along the DNA sequences (in the following the heterogeneity is called the 'segmentation effects'). At the first stage, the problem consists in recognizing the different elements of such a mosaic structure. Next, it is necessary to find whether or not some

Table 1. Characteristic lengths in genomic DNA sequences (in base pairs for double-stranded molecules and in bases for single-stranded molecules).

Mechanism	Length, l
The structure of a DNA helix	
Alteration of purines and pyrimidines in Z form	2
Helix pitch for B form	10.2 – 10.5
Helix pitch for A form	10.8
Helix pitch for Z form	12.0
Chromatin compactization	
Nucleosome	200 ± 40
Helix pitch for a 30 nm fibril	$(200 \pm 40) \times 6$
Loops	$2 \times 10^4 - 10^5$
Sub-unit composed of loops	~ 10 loops
Bending ability	
Kuhn length for double-stranded DNA	300 – 400
Kuhn length for single-stranded DNA	12 – 14
Protein structure	
Codon (corresponds to one amino acid)	3
Pitch for an α-helix	$(3.6 \pm 0.2) \times 3$
Alternation of hydrogen bonds in a β-sheet	2.0×3
Lengths of elements of a secondary structure	$(8 - 20) \times 3$
Protein lengths	$(70 - 2000) \times 3$
Intervening structure of protein-coding regions for eukaryots	
Exons	50 – 400
Introns	$50 - 5 \times 10^4$
Packing of heteronuclear RNA into RNP particles	$600 + 100$
The total lengths of genomes	
DNA viruses	$5 \times 10^3 - 5 \times 10^5$
Bacteria	$7 \times 10^5 - 10^7$
Plants	$10^8 - 10^{11}$
Amphibia	$6 \times 10^8 - 10^{11}$
Birds	$7 \times 10^8 - 2 \times 10^9$
Mammals	$(2 - 3) \times 10^9$

coupling exists between different elements of the mosaic structure, and if it exists, what molecular mechanisms are responsible for it. In Section 2.2 we briefly describe the main methods for analysis of DNA sequences.

2.2 Methods for statistical analysis of genomic DNA sequences

First of all let us settle a terminology. The lengths $l < 10$ we consider as small, the lengths $10 \leqslant l < 10^3$ as middle, and the lengths $l \geqslant 10^3$ as large. The length $l \sim 10$ corresponds to the pitch of a DNA double helix, $l \sim 10^3$ is the mean size of a gene, this length covers several nucleosomes, etc. (see Table 1). One of the natural scales is also the total length of a genomic sequence. This classification is used in Section 5. The terms 'an order' and 'ordering' are used in a sense which is traditional for physics, i.e., the terms reflect the existence of statistically significant correlations on given scales. In this sense 'the long-range order' includes the large-scale density variations, non-damping periodic oscillations, and tandem repeating of long random sequences as well.

The simplest approach to the statistical analysis of DNA sequences consists in the calculation of occurrence numbers or frequencies for different nucleotide combinations, $\{\mathrm{N}_1 \ldots \mathrm{N}_l\}$, $\mathrm{N} \in (\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T})$, when the sequence is subdivided into overlapping fragments of length l [5–9, 41–45] (in information theory such combinations are called l-grams [46]). For local analysis, the frequencies are calculated for combinations within a window which is of width W and slides along a sequence; then, the frequencies are processed with the help of the statistical tables. Since for four nucleotides the number of different l-grams is equal to 4^l and grows exponentially with the length l, in applications the length l is usually chosen to be rather small, $l \leqslant 6$ ($4^6 = 4096$).

If the occurrence numbers of nucleotide combinations of length l are divided by the total number of fragments in a sequence (for a sequence of length M this number equals $M - l + 1$), such frequencies can be considered as probabilities $P(\mathrm{N}_1 \ldots \mathrm{N}_l)$ and analyzed in the framework of a Markov chain formalism [5–7, 47–50] with the use of the well-known formula in probability theory [51]:

$$P(\mathrm{N}_1 \ldots \mathrm{N}_{l-1} | \mathrm{N}_l) = \frac{P(\mathrm{N}_1 \ldots \mathrm{N}_l)}{P(\mathrm{N}_1 \ldots \mathrm{N}_{l-1})}, \tag{2.1}$$

where $P(\mathrm{N}_1 \ldots \mathrm{N}_{l-1} | \mathrm{N}_l)$ is the conditional probability of the event such that the nucleotides $\mathrm{N}_1 \ldots \mathrm{N}_{l-1}$ are followed by N_l, $P(\mathrm{N}_1 \ldots \mathrm{N}_l)$ and $P(\mathrm{N}_1 \ldots \mathrm{N}_{l-1})$ are the probabilities for combinations $\mathrm{N}_1 \ldots \mathrm{N}_l$ and $\mathrm{N}_1 \ldots \mathrm{N}_{l-1}$, respectively. If there exists a number r such that the probability (2.1) does not depend on r for $l - 1 \geqslant r$, we can talk about a Markov chain of order r.

In terms of information theory [46], using the probabilities $P(\mathrm{N}_1 \ldots \mathrm{N}_l)$, we can calculate an entropy of order l,

$$H_l = - \sum_{\{\mathrm{N}_1 \ldots \mathrm{N}_l\}} P(\mathrm{N}_1 \ldots \mathrm{N}_l) \log_2 P(\mathrm{N}_1 \ldots \mathrm{N}_l), \tag{2.2}$$

where the summation is performed over all l-grams, a specific entropy

$$G_l = \frac{H_l}{l} \tag{2.3}$$

and a redundancy

$$R_l = 1 - \frac{H_l}{2l}. \tag{2.4}$$

The redundancy R_l characterizes the deviations from the completely random sequences in which all nucleotides occur with the same probability. Examples of calculations of these quantities can be found in Refs [5, 52–57].

The approaches based on both Markov chain and information theory can only be used provided the inequality $M - l + 1 \gg 4^l$ holds; in addition, both approaches implicitly assume a statistical homogeneity along a sequence. However, in the real genomic sequences there are segmentation effects which become pronounced at $M \geqslant 10^2 - 10^3$. Because of these effects, the approaches are applicable only for sufficiently small l (see also a critical analysis of the problem in Refs [54, 56]). Thus, the methods may only reveal small-scale correlations.

Middle-scale correlations can be analyzed with the help of correlation functions [17, 18, 23, 58, 59] or mutual information functions [60, 61]. These functions are defined in Section 3, therefore, we do not dwell on them here.

When studying large-scale density variations, the binary subdivision on purines $\mathrm{R} = (\mathrm{A}, \mathrm{G})$ and pyrimidines $\mathrm{Y} = (\mathrm{C}, \mathrm{T})$ is most commonly used. Further, a DNA sequence is split into segments of length l and the root-mean-square difference between the purine and pyrimidine numbers, $\langle [N_{\mathrm{R}}(l) - N_{\mathrm{Y}}(l)]^2 \rangle^{1/2}$, are calculated for different l [62–71]. If the purines and pyrimidines are considered to correspond to the numbers +1 and −1, respectively, such a model can be mapped onto the model of a random walk. In Ref. [62] the hypothesis was stated that such a random walk is fractal. In this case $\left\langle [N_{\mathrm{R}}(l) - N_{\mathrm{Y}}(l)]^2 \right\rangle^{1/2} \propto l^{\alpha}$, where α is a constant exponent different from 0.5. Now this hypothesis is rejected by most of scientists, because α is shown to depend on l [19, 63–65]. It is essential that α varies considerably when a sequence is split into several large segments and then α is calculated for each separate segment [65, 69]. In Ref. [66] it was shown that in exon-intron sequences the effective value of α is little affected by random rearrangements of introns. Thus, even though the difference between α and 0.5 is considerable, the question remains of a character of the large-scale correlations observed. In Refs [64–66] a model is suggested where a large-scale mosaic structure is considered, but the correlations between the elements of the mosaic are absent. However, in Refs [20, 27] it was shown that a non-trivial structural coupling exists between the elements (see also Section 5). The large-scale segmentation in DNA sequences can be investigated with the use of a technique with sliding windows [5] or wavelet transformation [72–76].

When chromosomal DNA is colored by dyes sensitive to S = (G, C) or W = (A, T), large-scale compositional variations with scales of $L \sim 10^5 - 10^6$ are observed [29, 77]. In Ref. [78] a model of a hidden random walk was suggested to explain the effects. To begin with, in such a model a random value of w is chosen within the interval $1/3 \leqslant w \leqslant 2/3$. Let at the initial moment of time some initial value w_0 be chosen. Then, w takes one of the values $w_0 - \Delta w$ or $w_0 + \Delta w$ with probabilities 1/2. On the boundaries of the interval the value of w remains constant. After m successive steps, the value of w will be equal to w_m, which, in turn, determines the probability $P(w_m)$ that the mth site will be occupied by the nucleotide W. The function $P(w)$ and the increment Δw are adjusted to match the experimental data. If the initial value w_0 is within the interval $1/2 < w_0 < 2/3$, then for approximately $L \sim [(w_0 - 1/2)/\Delta w]^2$ steps an increased probability for sites to be occupied by nucleotides W persists. In the interval $1/3 < w_0 < 1/2$ the opposite situation will be observed. In this model the mean size of regions with increased content of W or S is approximately equal to $\langle L \rangle \approx [(1/3 - 1/2)/\Delta w]^2$, however, the density variations within different regions of length $\sim L$ are uncorrelated. Detailed analysis of these correlations has not been carried out yet.

As is seen, the analysis of correlations at different scales is carried out by essentially different methods. Therefore, there is a problem concerning the development of a universal technique suitable for analysis at all scales. In the following sections we show that such a problem can be solved within the framework of a spectral approach.

3. Spectral analysis of DNA sequences

Spectral methods for analysis of DNA sequences are commonly used to reveal hidden periodicities [21, 26, 79–85], to study the correlations between different sequences [23, 27, 86–89], and to investigate long-range correlations [19, 27, 60, 90]. A general introduction to these methods can be found in Ref. [91]. The use of the Fourier transform allows one to obtain statistical criteria that already possess a self-averaging property for relatively short sequences of length $M \geqslant 100-200$. Beginning from these lengths, a segmentation in the genomic DNA sequences is observed (see Table 1). Therefore, within the framework of such a technique we can study both the separate elements of mosaic structure and the coupling between the elements. The strategy is to compare the observed characteristics for real sequences with that for random sequences with the same nucleotide composition. The last item is important when revealing segmentation effects as well as the effects of the compositional variations arising in the evolution and in consequence of the differences in the environmental conditions for different organisms. We show in the following that in this way one can obtain quite convenient and universal criteria for structural regularity. The criteria do not depend on the nucleotide composition thereby allowing to compare the structural characteristics of DNA sequences with the different nucleotide compositions. For the most part, in this section our exposition follows [17, 18].

3.1 General theory

Consider a DNA sequence of length M. It can be described in terms of a position function:

$$\rho_{m,\alpha} = \begin{cases} 1, & \text{if a nucleotide of a type } \alpha \text{ occupies the } m\text{th site}, \\ 0 & \text{otherwise}, \end{cases} \tag{3.1}$$

where $\alpha \in (\mathrm{A, C, G, T})$, $m = 1, \ldots, M$. The Fourier harmonics corresponding to nucleotides of type α are defined as

$$\rho_\alpha(q_n) = M^{-1/2} \sum_{m=1}^{M} \rho_{m,\alpha} \exp(-\mathrm{i} q_n m)\,,$$
$$q_n = \frac{2\pi n}{M}\,, \qquad n = 0, 1, \ldots, M-1\,. \tag{3.2}$$

The inverse transformation is given by

$$\rho_{m,\alpha} = M^{-1/2} \sum_{n=0}^{M-1} \rho_\alpha(q_n) \exp(\mathrm{i} q_n m)\,, \qquad m = 1, \ldots, M\,. \tag{3.3}$$

The zeroth Fourier harmonic does not contain any positional information and is determined only by the total number of nucleotides, N_α,

$$\rho_\alpha(0) = \frac{N_\alpha}{M^{1/2}}\,. \tag{3.4}$$

In the following the main characteristics are expressed in terms of the elements of a matrix structure factor,

$$F_{\alpha\beta}(q_n) = \rho_\alpha(q_n)\, \rho^*_\beta(q_n) \tag{3.5}$$

(henceforth the asterisk denotes complex conjugation). Using the Wiener–Khinchin relationship, one can relate the elements of the structure factors $F_{\alpha\beta}(q_n)$ with that of the pair correlation functions,

$$K_{\alpha\beta}(m_0) = M^{-1} \sum_{n=0}^{M-1} F_{\alpha\beta}(q_n) \exp(-\mathrm{i} q_n m_0)\,,$$
$$m_0 = 0, \ldots, M-1\,. \tag{3.6}$$

Using definitions (3.2) and (3.3), the correlation functions can be written as

$$K_{\alpha\beta}(m_0) = M^{-1} \sum_{m=1}^{M} \rho^{\mathrm{c}}_{m,\alpha}\, \rho^{\mathrm{c}}_{m+m_0,\beta}\,, \tag{3.7}$$

where

$$\rho^{\mathrm{c}}_{m,\alpha} = \begin{cases} \rho_{m,\alpha}\,, & \text{if } \ 1 \leqslant m \leqslant M\,, \\ \rho_{m-M,\alpha}\,, & \text{if } \ M+1 \leqslant m \leqslant 2M-1\,. \end{cases} \tag{3.8}$$

Since the functions $\rho_{m,\alpha}$ are real-valued, it follows that

$$\rho^*_\alpha(q_n) = \rho_\alpha(2\pi - q_n)\,. \tag{3.9}$$

This relationship, in its turn, imposes the following symmetry conditions on the elements of the structure factors and correlation functions:

$$F_{\alpha\beta}(q_n) = F_{\beta\alpha}(2\pi - q_n)\,,$$

$$K_{\alpha\beta}(m_0) = K_{\beta\alpha}(M - m_0)\,. \tag{3.10}$$

If $a = \beta$, conditions (3.10) allow one to take into consideration only the left half-spectra, $1 \leqslant n \leqslant N$, $1 \leqslant m_0 \leqslant N$, where

$$N = \left[\frac{M}{2}\right], \tag{3.11}$$

and the square brackets denote the integral part of the number.

Since each position in a sequence is occupied by only one nucleotide, we have the equalities

$$\sum_\alpha \rho_{m,\alpha} = 1\,, \tag{3.12}$$

$$\sum_\alpha \rho_\alpha(q_n) = 0 \qquad (n \neq 0)\,. \tag{3.13}$$

These restrictions express the excluded-volume effects and result in non-trivial correlations even for random sequences. From (3.13) it follows that $F_{11}(q_n) = F_{22}(q_n)$ for binary sequences.

An important property of the theory is the existence of a number of exact sum rules. Below we make use of the following sum rules:

$$\bar{F}_{\alpha\beta} = (M-1)^{-1} \sum_{n=1}^{M-1} F_{\alpha\beta}(q_n) = \frac{\delta_{\alpha\beta} N_\alpha - N_\alpha N_\beta / M}{M-1}\,, \tag{3.14}$$

$$\bar{K}_{\alpha\beta} = (M-1)^{-1} \sum_{m_0=1}^{M-1} K_{\alpha\beta}(m_0) = \frac{N_\alpha N_\beta - \delta_{\alpha\beta} N_\beta}{M(M-1)}\,, \tag{3.15}$$

$$\sum_{m_0=1}^{M-1} \left[K_{\alpha\beta}(m_0) - \bar{K}_{\alpha\beta}\right]\left[K_{\gamma\delta}(m_0) - \bar{K}_{\gamma\delta}\right] = M^{-1} \sum_{n=1}^{M-1} \left[F_{\alpha\beta}(q_n) - \bar{F}_{\alpha\beta}\right]\left[F^*_{\gamma\delta}(q_n) - \bar{F}^*_{\gamma\delta}\right], \tag{3.16}$$

where $\delta_{\alpha\beta}$ is the Kronecker delta. As is seen from (3.14) and (3.15), for sequences with the fixed nucleotide composition the mean characteristics $\bar{F}_{\alpha\beta}$ and $\bar{K}_{\alpha\beta}$ are identical.

Because the Fourier transform is completely reversible, each structural feature of a DNA sequence corresponds to some counterpart in the spectra for structure factors and correlation functions. For example, a periodicity results in the periodic variations in $K(m_0)$ and a series of equidistant peaks for $F(q_n)$ (see Section 4). As another example, we consider an elementary variant of segmentation when adenines occupy sites from 1 up to L, while the other nucleotides are located beyond this region. Then we obtain

$$F_{\mathrm{AA}}(q_n) = \frac{\sin^2(q_n L/2)}{M \sin^2(q_n/2)}\,, \tag{3.17}$$

where a characteristic growth is seen in the region of small wave numbers, $q_n \leqslant 1/L$. For correlation functions with $\alpha = \beta$ and finite correlation radius r_{c} we have

$$\Delta K(m_0) = K(m_0) - \bar{K} \propto \exp\left(-\frac{m_0}{r_{\mathrm{c}}}\right) + \exp\left(-\frac{M - m_0}{r_{\mathrm{c}}}\right), \tag{3.18}$$

hence,

$$\Delta F(q_n) \propto \left[1 - \exp\left(-\frac{M}{r_{\mathrm{c}}}\right)\right] \frac{\cos q_n - \exp(-1/r_{\mathrm{c}})}{\cosh(1/r_{\mathrm{c}}) - \cos q_n}\,. \tag{3.19}$$

The corresponding half-spectra for $M = 100$ are shown in Fig. 2 for a case of correlations. When anticorrelations arise, it is necessary to change the signs of ΔK and ΔF.

In genome formation, important roles are played by both the direct duplication of DNA fragments and the duplication of inverted and complementary fragments [4, 41]. Let the Fourier harmonics for an initial sequence be determined by (3.2). Then to obtain the inverted sequence, we have to change the beginning and the end, hence [21],

$$\rho^{\mathrm{I}}_\alpha(q_n) = \exp(-\mathrm{i} q_n)\,\rho_\alpha(2\pi - q_n)\,, \tag{3.20}$$

where $\rho^{\mathrm{I}}_\alpha(q_n)$ are the Fourier harmonics for the inverted sequence. If a sequence is composed of direct and inverted fragments, then it is invariant with respect to inversion. In this

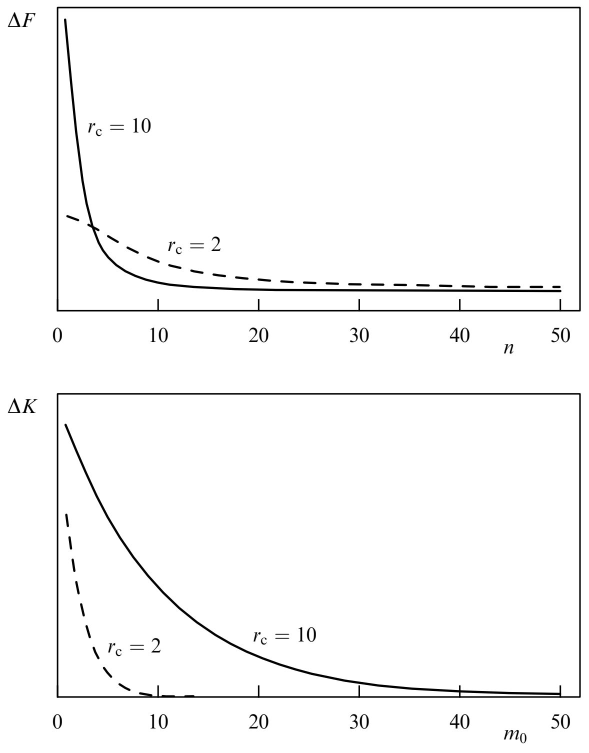

Figure 2. Dependences for the deviations of Fourier harmonic amplitudes ΔF and correlation functions ΔK from their mean values for different radii of correlations.

case it immediately follows from (3.20) that $\rho_\alpha(\pi) = 0$. The complementary sequence is obtained from an initial sequence by inversion and the replacements A ↔ T, G ↔ C. Therefore, for a sequence composed of direct and complementary fragments, the following equalities hold:

$$F_{\mathrm{AA}}(q_n) = F_{\mathrm{TT}}(q_n)\,, \qquad F_{\mathrm{CC}}(q_n) = F_{\mathrm{GG}}(q_n)\,. \tag{3.21}$$

Thus, the spectral analysis allows, in principle, to reveal all the structural features of the analyzed DNA sequence.

3.2 Statistical characteristics for random sequences

In the genomic DNA sequences, all regular features are observed on a strong random background caused by point mutations (this is the main source of random modifications [4, 29]), insertions, deletions, translocations, etc. Therefore, the statistical significance of the regular features observed should be estimated with respect to the corresponding characteristics for random DNA sequences with the same nucleotide composition.

The statistical distribution for Fourier harmonics can be obtained by averaging the characteristic function (see, e.g., Refs [51, 92])

$$Z = \exp\left[\mathrm{i}\sum_\alpha \sum_{n=1}^{M-1} u_\alpha(q_n)\,\rho_\alpha(q_n)\right] \tag{3.22}$$

over an ensemble of random realizations of sequences with fixed total numbers of nucleotides $\{N_\alpha\}$. It is convenient to consider the auxiliary variables $u_\alpha(q_n)$ to obey the conditions

$$u^*_\alpha(q_n) = u_\alpha(2\pi - q_n)\,, \tag{3.23}$$

which are analogous to (3.9). The different products of Fourier harmonics are obtained by the differentiation of Z with respect to the auxiliary variables $u_\alpha(q_n)$ with subsequent equating $u_\alpha(q_n)$ to zero, e.g.,

$$\rho_\alpha(q_n) = \left.\frac{\partial Z}{\mathrm{i}\partial u_\alpha(q_n)}\right|_{u_\alpha=0}, \tag{3.24}$$

etc. Using definitions (3.1) and (3.2), we rewrite Z in the form

$$Z = \prod_\alpha \prod_{m=1}^{M} (1 + \rho_{m,\alpha}\, z_{m,\alpha})\,, \tag{3.25}$$

$$z_{m,\alpha} = \exp\left[\mathrm{i}M^{-1/2}\sum_{n=1}^{M-1} u_\alpha(q_n)\exp(-\mathrm{i}q_n m)\right] - 1\,. \tag{3.26}$$

Thus, the problem is reduced to averaging the different products of $\rho_{m,\alpha}$.

The averaging is performed with the use of simple combinatorial considerations and the result is given by

$$\left\langle \prod_{k_1=1}^{n_\mathrm{A}} \rho_{m_{k_1},\mathrm{A}} \prod_{k_2=1}^{n_\mathrm{C}} \rho_{m_{k_2},\mathrm{C}} \prod_{k_3=1}^{n_\mathrm{G}} \rho_{m_{k_3},\mathrm{G}} \prod_{k_4=1}^{n_\mathrm{T}} \rho_{m_{k_4},\mathrm{T}} \right\rangle = \frac{C^{M-n_\mathrm{A}-n_\mathrm{C}-n_\mathrm{G}-n_\mathrm{T}}_{N_\mathrm{A}-n_\mathrm{A},\,N_\mathrm{C}-n_\mathrm{C},\,N_\mathrm{G}-n_\mathrm{G},\,N_\mathrm{T}-n_\mathrm{T}}}{C^M_{N_\mathrm{A},\,N_\mathrm{C},\,N_\mathrm{G},\,N_\mathrm{T}}}\,, \tag{3.27}$$

$$C^m_{n_\mathrm{A},\,n_\mathrm{C},\,n_\mathrm{G},\,n_\mathrm{T}} = \frac{m!}{n_\mathrm{A}!\,n_\mathrm{C}!\,n_\mathrm{G}!\,n_\mathrm{T}!}\,,$$

$$n_\mathrm{A} + n_\mathrm{C} + n_\mathrm{G} + n_\mathrm{T} = m\,, \qquad 0! = 1\,. \tag{3.28}$$

Here the angular brackets denote averaging over the ensemble of random realizations, all the subscripts $\{m_{k_1}\}, \ldots, \{m_{k_4}\}$ are assumed to be different due to excluded-volume effects, n_α is the total number of position functions for nucleotides of type α, and N_α is the total number of nucleotides of type α in sequences of length M. The right-hand side in (3.27) is equal to the ratio of two combinatorial factors, i.e., $C^M_{N_\mathrm{A}, N_\mathrm{C}, N_\mathrm{G}, N_\mathrm{T}}$, which is the total number of different random realizations, and

$$C^{M-n_\mathrm{A}-n_\mathrm{C}-n_\mathrm{G}-n_\mathrm{T}}_{N_\mathrm{A}-n_\mathrm{A},\,N_\mathrm{C}-n_\mathrm{C},\,N_\mathrm{G}-n_\mathrm{G},\,N_\mathrm{T}-n_\mathrm{T}}\,,$$

which is equal to the total number of realizations provided that n_A positions are occupied by nucleotides A, and analogously for n_C, n_G and n_T [it is the positions that enter the left-hand side in (3.27)].

Straightforward calculation of the lowest-order averages with $q_n \neq 0$ gives:

$$\langle \rho_\alpha(q_n) \rangle = 0\,, \tag{3.29}$$

$$\langle \rho_\alpha(q_n)\rho_\alpha(q_{n'}) \rangle = \begin{cases} \bar{F}_{\alpha\beta}\,, & q_n + q_{n'} = 2\pi\,, \\ 0, & q_n + q_{n'} \neq 2\pi\,, \end{cases} \tag{3.30}$$

where $\bar{F}_{\alpha\beta}$ is defined by (3.14). Equality (3.30) reflects an important quasi-ergodic property, namely, averaging over an ensemble is asymptotically equivalent to averaging over a spectrum. Provided that $M \gg 1$, $N_\mathrm{A} \gg 1, \ldots, N_\mathrm{T} \gg 1$, and taking into account (3.23), we can prove that the following expression for $\langle Z \rangle$ can be used in the main approximation with respect to $M^{-1/2}$ [17, 18]:

$$\langle Z \rangle \approx \exp\left[-\sum_{\alpha,\beta}\sum_{n=1}^{N} \bar{F}_{\alpha\beta}\, u_\alpha(q_n)\, u^*_\beta(q_n)\right], \tag{3.31}$$

where N is given by (3.11). Using (3.31), we can obtain a number of useful particular criteria for regularity.

Mutual correlations. The correlations in the positions of different nucleotides are characterized in terms of a cross correlation coefficient [51, 92],

$$k(F_{\alpha\beta}|F_{\gamma\delta};\, M-1) = \sum_{n=1}^{M-1} \frac{\left[F_{\alpha\beta}(q_n) - \bar{F}_{\alpha\beta}\right]\left[F^*_{\gamma\delta}(q_n) - \bar{F}^*_{\gamma\delta}\right]}{(M-1)\,\sigma(F_{\alpha\beta})\,\sigma(F_{\gamma\delta})}\,, \tag{3.32}$$

$$\sigma^2(F_{\alpha\beta};\, M-1) = \sum_{n=1}^{M-1} \frac{\left[F_{\alpha\beta}(q_n) - \bar{F}_{\alpha\beta}\right]\left[F^*_{\alpha\beta}(q_n) - \bar{F}^*_{\alpha\beta}\right]}{M-1}\,. \tag{3.33}$$

If k approaches unity, then the nucleotides become completely correlated, while if k vanishes, then correlations are absent. Similar characteristics can be defined for correlation functions (3.7) as well, however, from the exact sum rule (3.16) it follows that

$$\sigma(K_{\alpha\beta};\, M-1) = \frac{\sigma(F_{\alpha\beta};\, M-1)}{M^{1/2}}\,, \tag{3.34}$$

$$k(K_{\alpha\beta}|K_{\gamma\delta};\, M-1) = k(F_{\alpha\beta}|F_{\gamma\delta};\, M-1)\,. \tag{3.35}$$

Using the asymptotic equivalence of averaging over the ensemble and over the spectrum, for random sequences we obtain

$$\langle k(F_{\alpha\beta}|F_{\gamma\delta};\, M-1) \rangle = \frac{\bar{F}_{\alpha\gamma}\bar{F}_{\delta\beta}}{(\bar{F}_{\alpha\alpha}\bar{F}_{\beta\beta}\bar{F}_{\gamma\gamma}\bar{F}_{\delta\delta})^{1/2}}\,, \tag{3.36}$$

in particular, if $\alpha \neq \beta$, we have

$$\big\langle k(F_{\alpha\alpha}|F_{\beta\beta};\, M-1)\big\rangle \equiv \langle k_{\alpha\beta}\rangle = \frac{N_\alpha N_\beta}{(M-N_\alpha)(M-N_\beta)}\,. \quad (3.37)$$

Equation (3.37) has a simple physical interpretation. The correlation coefficient $\langle k(F_{\alpha\alpha}|F_{\beta\beta};\, M-1)\rangle$ is equal to the probability of simultaneously finding nucleotides of type α in the positions free from nucleotides of type β and vice versa. It is seen that the excluded-volume effects result in non-trivial correlations between different nucleotides even in random sequences. The root-mean-square deviations from the average value of the correlation coefficient can be estimated as [93]

$$\Big\langle \big[\Delta k(F_{\alpha\alpha}|F_{\beta\beta};\, M-1)\big]^2\Big\rangle \equiv \big\langle (\Delta k_{\alpha\beta})^2\big\rangle \approx \frac{1}{N}\,. \quad (3.38)$$

We do not present here the correction terms for (3.38) due to the finite value of the correlation coefficient [23]. It is essential that these corrections are negative, hence, (3.38) can only strengthen the estimate for the statistical significance of the normalized deviation,

$$\frac{\big[k(F_{\alpha\alpha}|F_{\beta\beta};\, M-1) - \langle k_{\alpha\beta}\rangle\big]}{\big[2\langle(\Delta k_{\alpha\beta})^2\rangle\big]^{1/2}}\,,$$

which has a Gaussian distribution for random sequences.

Distribution of harmonic amplitudes. Taking into account that the characteristic function and multivariate distribution function are related to each other by the Fourier transform [51, 92] and using (3.31), we can obtain a distribution function for harmonic amplitudes in the random spectra [the amplitudes are equal to the diagonal elements of the structure factor (3.5)]. The probability that the amplitude of the nth harmonic lies between $F_{\alpha\alpha}(q_n)$ and $F_{\alpha\alpha}(q_n) + \mathrm{d}F_{\alpha\alpha}(q_n)$ is

$$p\big(F_{\alpha\alpha}(q_n)\big)\,\mathrm{d}F_{\alpha\alpha}(q_n) = \exp(-f_{n,\alpha\alpha})\,\mathrm{d}f_{n,\alpha\alpha}\,, \quad (3.39)$$

$$f_{n,\alpha\alpha} = \frac{F_{\alpha\alpha}(q_n)}{\bar{F}_{\alpha\alpha}}\,. \quad (3.40)$$

Distribution function (3.39) is the Rayleigh distribution [51], which is every bit as universal in spectral analysis as is the Gaussian distribution for real random variables.

As is seen from (3.39), the probability that some harmonic exceeds a value $F_{\alpha\alpha}$ is given by

$$P\big\{F_{\alpha\alpha}(q_n) > F_{\alpha\alpha}\big\} = \int_{F_{\alpha\alpha}}^{\infty} p(F'_{\alpha\alpha})\,\mathrm{d}F'_{\alpha\alpha} = \exp\left(-\frac{F_{\alpha\alpha}}{\bar{F}_{\alpha\alpha}}\right). \quad (3.41)$$

This also means that in the half-spectrum the average number of harmonics with amplitudes exceeding $F_{\alpha\alpha}$ is equal to

$$\langle n_\alpha\rangle = N\exp\left(-\frac{F_{\alpha\alpha}}{\bar{F}_{\alpha\alpha}}\right) \quad (3.42)$$

[here we take into account that only N harmonics are statistically independent in accordance with (3.10) and (3.11)]. The condition $\langle n_\alpha\rangle = 1$ determines the characteristic value of amplitudes for singular outbursts in the random spectra,

$$F_{\alpha\alpha,\,\max} \approx \bar{F}_{\alpha\alpha}\ln N\,. \quad (3.43)$$

Structural entropy of a sequence. The cross correlation coefficients (3.32) characterize only the mutual positions of nucleotides but not the ordering of a sequence. Indeed, two almost identical random sequences are strongly correlated while remaining random. The integral ordering of a sequence can be estimated in terms of the structural entropy

$$S_\alpha = -\sum_{n=1}^{M-1} f_{n,\alpha\alpha}\ln f_{n,\alpha\alpha}\,, \quad (3.44)$$

where $f_{n,\alpha\alpha}$ are defined by (3.40). Taking into account that $\{f_{n,\alpha\alpha}\}$ obey the sum rule (3.14),

$$\sum_{n=1}^{M-1} f_{n,\alpha\alpha} = \mathrm{const}\,, \quad (3.45)$$

it is easily seen that under the condition (3.45) S_α attains a maximum when the distribution of amplitudes in the spectrum is strictly uniform. Since the heights of Fourier harmonics are distributed over a spectrum more uniformly for random sequences than for regular ones (see below), we see that S_α may serve as an approximate measure of ordering.

Averaging S_α with the use of the probability distribution function (3.39), we obtain the mean structural entropy for random sequences:

$$\langle S_\alpha\rangle_{\text{random}} = -(1-C)(M-1)\,, \quad (3.46)$$

where $C = 0.577215\ldots$ is the Euler constant. To determine the standard deviations, we have to take into account the correlation between harmonics with different q_n; calculation gives $\big\langle(\Delta S_\alpha)^2\big\rangle_{\text{random}} \simeq (0.5797\ldots)(M-1)$.

The total structural entropy is given by

$$S = \sum_\alpha S_\alpha\,. \quad (3.47)$$

To compare the regularity of sequences of different lengths, it is convenient to introduce the relative structural entropy $\Delta S_{\text{rel}} = (\langle S\rangle_{\text{random}} - S)/|\langle S\rangle_{\text{random}}|$, because for regular deviations from randomness we have $\langle S\rangle_{\text{random}} - S \propto (M-1)$, hence, ΔS_{rel} does not depend on M. For random sequences $\Delta S_{\text{rel}} \propto (M-1)^{-1/2}$, this value is small for sufficiently large M. Since ΔS_{rel} is approximately independent of sequence length and nucleotide composition, it follows that this characteristic can be used as a universal measure of structural regularity for various sequences. Examples of applications of relative structural entropy can be found in Refs [19, 24].

3.3 Examples of spectra for genomic DNA sequences

As an illustration, we consider the structure factor spectra for the genomic DNA sequence of the bacteriophage PHIX174 (the accession numbers in EMBL database are V01128 and J02482). On the one hand, the molecular-biological structure of this genome is well-studied [29, 94, 95], and on the other hand, the genomic sequence possesses some universal features, which are discussed here and in the following sections. The genome of PHIX174 is represented by a single-stranded circular DNA of length $M = 5386$ with the nucleotide composition $N_{\mathrm{A}} = 1291$, $N_{\mathrm{C}} = 1157$, $N_{\mathrm{G}} = 1254$, $N_{\mathrm{T}} = 1684$. The genome is composed of 11 different genes, two of which are compound and some others are partially overlapping.

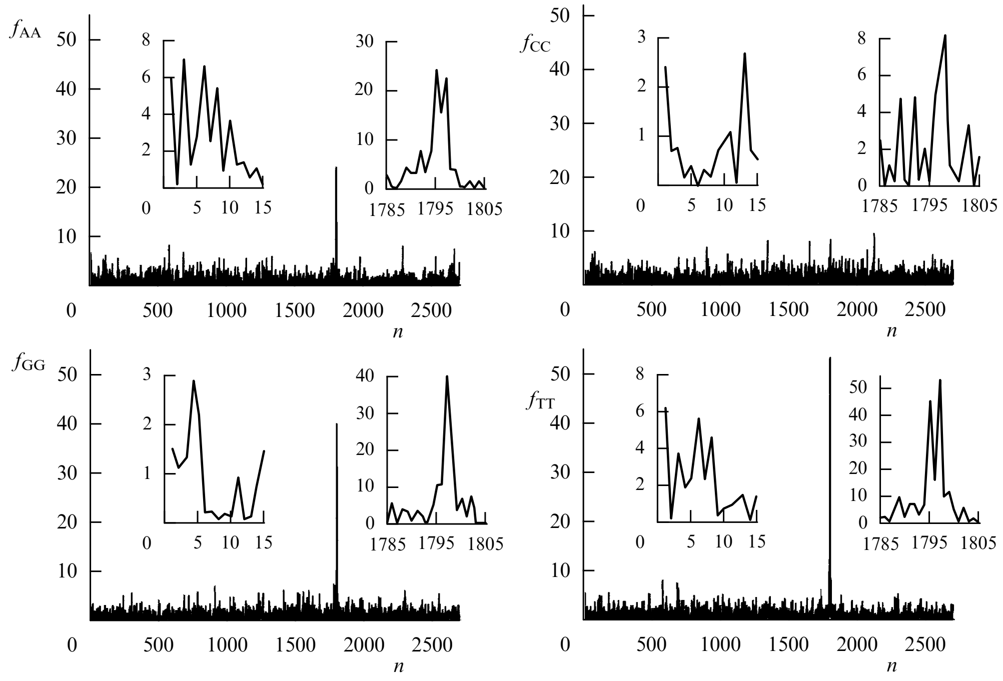

Figure 3. Half-spectra for the normalized diagonal elements of the structure factor [see (3.5) and (3.40)] calculated for the genomic DNA of the bacteriophage PHIX174.

In the following we consider only the diagonal elements of the structure factor (3.5). The corresponding half-spectra for the normalized harmonics (3.40) are shown in Fig. 3. All the Fourier spectra are plotted against the number n of wave numbers $q_n = 2\pi n/M$ for reasons which are explained in Section 4. To calculate the formal periods, we can use the relationship $p = M/n$. The inserts in Fig. 3 show the regions of small wave numbers, where we can observe the effects of large-scale segmentation [see (3.17)], and the vicinity of high peaks corresponding to periodicity $p = 3$, which is a universal feature of protein-coding regions (see Section 4.3). These peaks have a small half-width of $\sim 1-2$ harmonics, indicating that the periodicity is not damping, and the peaks for f_{AA} and f_{TT} have a characteristic split structure.

Below we use the following result obtained in Ref. [18]. Let $k = 2, 3, \ldots$ be a sequence of integers. If n/M is in the vicinity of $1/k$ and m runs from 1 to M, then the number of modulations in the envelope of the maxima of $\cos(q_n m)$ (where $q_n = 2\pi n/M$) is equal to

$$n_s = |kn - M| . \tag{3.48}$$

The transition from the period k to $k+1$ occurs at n such that

$$(k+1)n - M = M - kn . \tag{3.49}$$

Modulational superperiods can be observed if within the interval $\big(2\pi/(k+1),\, 2\pi/k\big)$ there exist at least two different wave numbers q_n, e.g., under the following condition:

$$\frac{1}{k} - \frac{1}{k+1} > \frac{1}{M} . \tag{3.50}$$

In particular, the highest harmonic $n = 1795$ for f_{AA} results in a unique modulational superperiod, and the highest harmonics $n = 1797$ for f_{GG} and f_{TT} generate five modulational superperiods.

Figure 4 shows the dependences for a fraction of harmonics $N(f)/N$ with the amplitudes exceeding a given threshold value f. It is seen in Fig. 4 that only slightly more than ten harmonics deviate from the exponential distribution (3.42) for random sequences.

For correlation functions (3.6), (3.7) it is convenient to use the variables

$$\varkappa_{\alpha\alpha}(m_0) = \frac{K_{\alpha\alpha}(m_0) - \bar{K}_{\alpha\alpha}}{\left[2\left\langle(\Delta K_{\alpha\alpha})^2\right\rangle\right]^{1/2}} , \tag{3.51}$$

where $\bar{K}_{\alpha\alpha}$ is given by (3.15) and

$$\left\langle(\Delta K_{\alpha\alpha})^2\right\rangle = \frac{\bar{F}^2_{\alpha\alpha}}{M} . \tag{3.52}$$

For random sequences, we can consider $\varkappa_{\alpha\alpha}(m_0)$ with different m_0 to be approximately independent Gaussian variables. The general view of the corresponding half-spectra is given in Fig. 5 and in more detail in Fig. 6. Oscillations with period $p = 3$ are clearly seen in the left panel of Fig. 6. The right panel of Fig. 6 shows the current standard deviations averaged over 100 sites,

$$\tilde{\sigma} = \left[100^{-1} \sum_{m_0'=m_0}^{m_0+99} \varkappa^2_{\alpha\alpha}(m_0)\right]^{1/2} , \tag{3.53}$$

where $m_0 = 1, 101, 201, \ldots, M/2$. In the region $m_0 \sim 500$ corresponding to the mean gene length, we observe a characteristic decrease in the standard deviations, however for T, large-scale variations with a scale comparable to the total genome length M are also clearly seen.

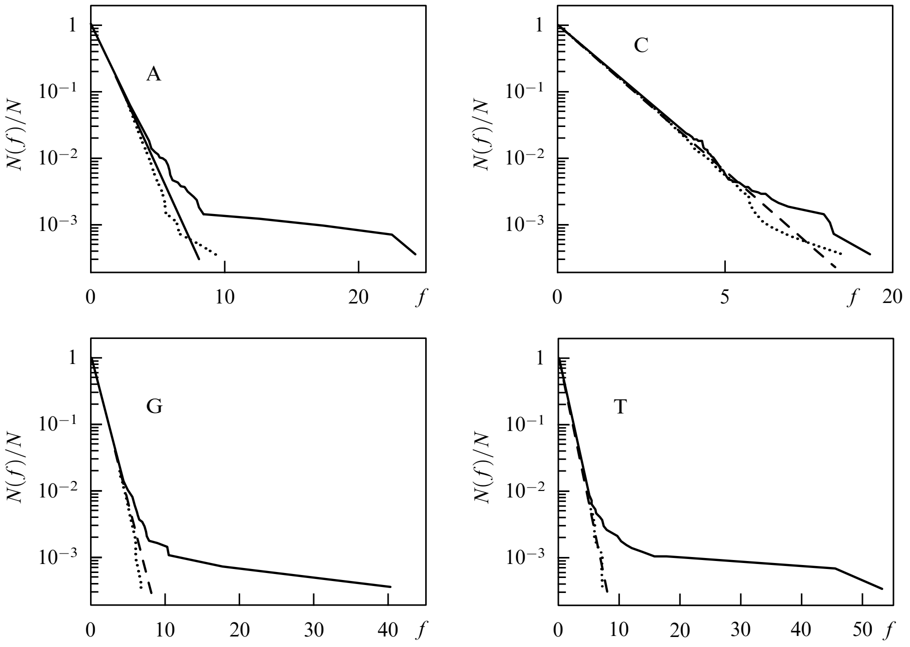

Figure 4. Dependences on f for the fraction of harmonics with an amplitude exceeding the value f. The dependence of the corresponding mean values for random sequences with the same nucleotide composition are shown by dashed lines [see (3.42)]. The dotted lines show the dependence for one of the random realizations.

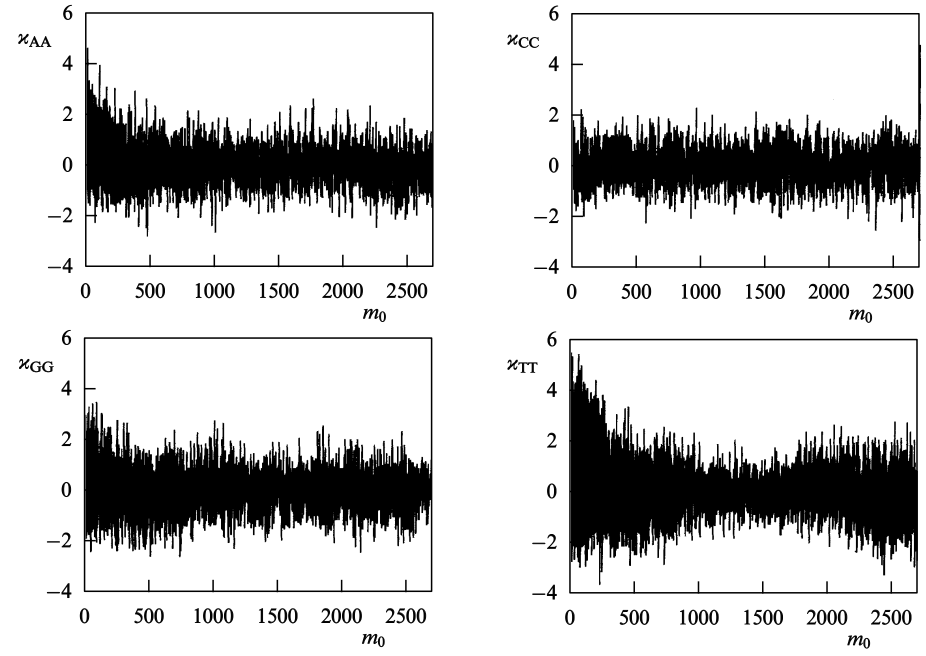

Figure 5. Half-spectra for the normalized correlation functions (3.51) calculated for the genomic DNA of the bacteriophage PHIX174.

The relative structural entropies (see the definition at the end of Section 3.2) for the PHIX174 genome are equal to

$$\Delta S_{\rm rel,\,A} = 1.41 \times 10^{-1}\,; \qquad \Delta S_{\rm rel,\,C} = -1.75 \times 10^{-3}\,;$$

$$\Delta S_{\rm rel,\,G} = 1.20 \times 10^{-1}\,; \qquad \Delta S_{\rm rel,\,T} = 2.25 \times 10^{-1}\,;$$

$$\Delta S_{\rm rel} = 1.28 \times 10^{-1}\,.$$

Only for A, G, and T do the deviations from mean random values appear to be statistically significant. The values of $\Delta S_{\rm rel}$ vary considerably for different genes [19].

This example shows that the structural characteristics for different nucleotides may strongly differ in genomic DNA sequences. Thus, we should always begin with the analysis for all four nucleotides. Only at the next stage can we combine different nucleotides to obtain the binary sequences.

4. Hidden periodicities in genomic DNA sequences

In Section 2.1 we have already pointed out the important role of repeated DNA fragments in genomes of higher organisms

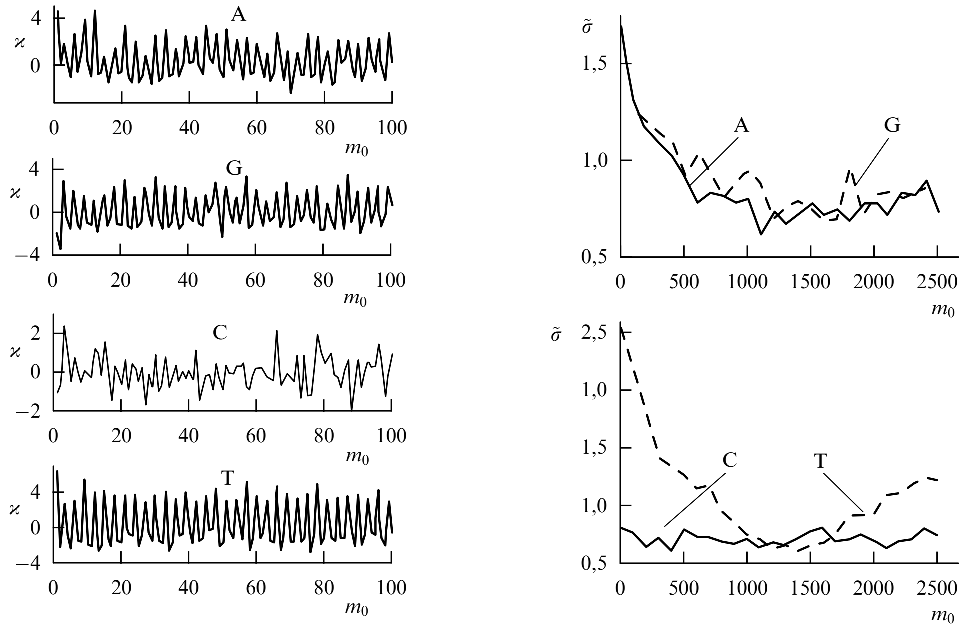

Figure 6. Initial part of the spectra shown in Fig. 5 (left panel). The smoothed standard deviations for the normalized correlation functions defined by (3.53) (right panel).

[29, 31, 35]. In the hidden form, the repeats modified by random mutations, insertions, deletions, etc. are typical for almost all genomic DNA sequences. In part, this is due to the periodicity in the double-stranded DNA structure, the approximately periodic structure of nucleosomes and 30-nm fibrils, etc. [38 – 40]. Some periodicities like the periodicity $p = 3$, which is inherent to the protein-coding regions [5, 41, 45], have, probably, an evolutionary origin related to the genetic code [96, 97]. Other periodicities in protein-coding regions can be related to the regularity and segmentation of protein secondary structure elements [98, 99]. Some applications of these ideas are outlined in Ref. [22]. Some periodicities can be explained by the repeating sites for cooperative binding of DNA with the regulatory proteins [29]. Finally, we mention the periodicities related to the evolutionary selection and subsequent duplication of some structures [4, 28, 29]. In this section we describe a general technique [21, 26], which is based on the spectral approach, for identifying hidden periodicities.

4.1 Structural features of hidden periodicities in genomic DNA sequences

To begin with, consider an idealized case where a fragment of length L is repeated N_p times, i.e., $M = LN_p$. The Fourier harmonics (3.2) for such a sequence can be written as

$$\rho_\alpha(q_n) = \frac{\rho_{\alpha,\,\mathrm{repeat}}(q_n)\left\{1 + \exp(-\mathrm{i}q_n L) + \ldots + \exp\left[-\mathrm{i}(N_p - 1)q_n L\right]\right\}}{N_p^{1/2}}\,, \tag{4.1}$$

$$\rho_{\alpha,\,\mathrm{repeat}}(q_n) = L^{-1/2}\sum_{m=1}^{L}\rho_{m,\alpha}\exp(-\mathrm{i}q_n L)\,,$$

$$q_n = \frac{2\pi n}{M}\,, \qquad n = 0, 1, \ldots, M-1\,. \tag{4.2}$$

Hence, for diagonal elements of the structure factor (3.5) we obtain

$$F_{\alpha\alpha}(q_n) = \frac{F_{\alpha\alpha,\,\mathrm{repeat}}(q_n)\sin^2(q_n L N_p/2)}{N_p\sin^2(q_n L/2)}\,. \tag{4.3}$$

It is easily seen from (4.3) that the repeats of length L result in a series of $L-1$ equidistant peaks at $q_n = \pi k/L$, $k = 1, \ldots, L-1$. This is due to the fact that in the general case a periodicity generates a series of equidistant peaks, the Fourier half-spectra are represented in terms of numbers n rather than the formal periods $p = M/n$.

The normalized half-spectra [see (3.40)] for 10^2 repeats ATAAACT in the genome of *Drosophila virilis* are shown in Fig. 7 (left panel). The right panel of the figure demonstrates the corresponding spectra for a sequence obtained after 45% random substitutions, with the probabilities being proportional to the contents of the corresponding nucleotides in the original sequence.

Now let in the sequence composed of 10^2 repeats ATAAACT these repeats be consecutively numerated and subdivided into even and odds. Then, let the nucleotides C be replaced by G in a randomly chosen even repeats. The replacements result in a characteristic modification of the spectra for nucleotides C, when between the original high equidistant peaks a series of equidistant peaks of lower height arise (Fig. 8). The procedure can be interpreted as a partial formation of the diblock complexes ATAAACTATAAAGT or a partial doubling of the initial period. The procedure can be continued, i.e., diblock 14-site units are numerated and subdivided into even and odds, and a fraction of even diblocks can then be modified analogously, etc. The procedure results in a characteristic hierarchical system of equidistant peaks and can be considered as a cascading period doubling. In the following such a process will be referred to as a multiplication of the initial period or a formation of structural subharmonics. The process appears

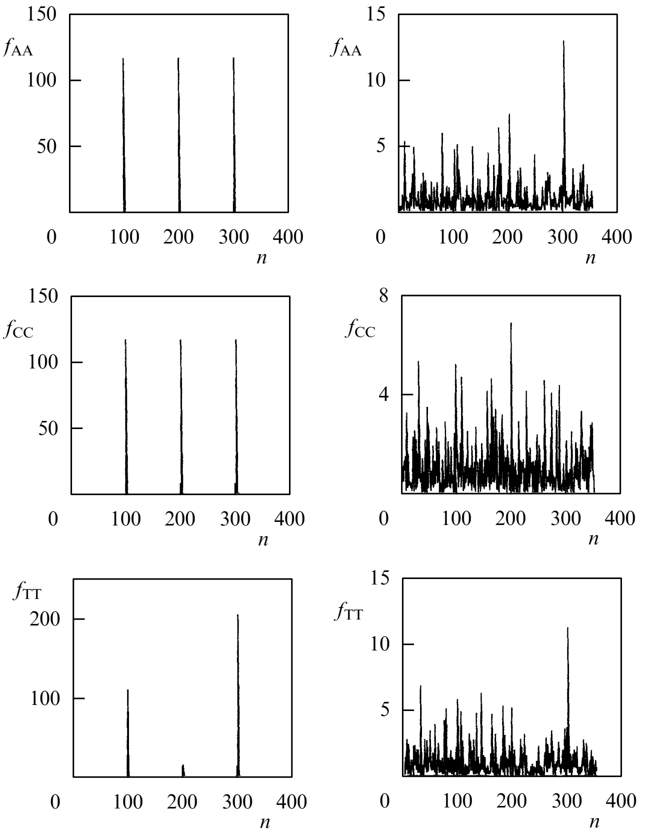

Figure 7. Half-spectra for the normalized harmonics for a sequence composed of 100 repeats ATAAACT (left panel). The same half-spectra after ~ 45% random substitutions with the probabilities proportional to the occurrences of the corresponding nucleotides in the original sequence (right panel).

to be characteristic for genomic DNA sequences. We can state that almost every well-pronounced period in the DNA sequences generates structural subharmonics. In general, the multiplication of the initial period is not necessarily doubling and is specific for different genomes. In addition, the peak positions allow one to determine the substitutions that are responsible for the multiplication of the period (compare the left and right panels of Fig. 8).

Another characteristic feature of genomic sequences is a coexistence of different periodicities and, as a consequence, the effects of mutual modulation. We explain the effect by the elementary example of density variations of the type

$$\Delta\rho_m \propto \cos\left(\frac{2\pi m}{P}\right)\cos\left(\frac{2\pi m}{p_0}\right), \tag{4.4}$$

where $P \gg p_0$. Such modulations result in splitting the period p_0 into two periods,

$$\frac{1}{p_{1,2}} = \frac{1}{p_0} \pm \frac{1}{P}\,. \tag{4.5}$$

Due to similar effects of mutual modulation, it is necessary to consider not only a series of strictly equidistant harmonics but also a series of harmonics with slightly violated equidistance, i.e., each harmonic in an equidistant series, $F_{\alpha\alpha}(q_n), \ldots, F_{\alpha\alpha}(rq_n)$, is surrounded by a window of width w,

$$F_{\alpha\alpha}(kq_n) = F_{\alpha\alpha}(q_{kn}) \\ \to F_{\alpha\alpha}(q_{kn-w}), \ldots, F_{\alpha\alpha}(q_{kn}), \ldots, F_{\alpha\alpha}(q_{kn+w})\,, \tag{4.6}$$

then, the harmonic with maximum height is chosen of all $2w+1$ harmonics in each window. The optimum choice of a window width w depends on the character of modulations.

The use of window w worsens the resolution of the period. Since

$$\frac{2\pi(n_p \pm w)}{M} = \frac{2\pi}{p \pm \Delta p} \approx \frac{2\pi}{p}\left(1 \mp \frac{\Delta p}{p}\right),$$

we obtain the estimate $\Delta p \sim 2wp^2/M$ for the period uncertainty. Thus, longer periods should be investigated with narrower windows. If we use windows beginning from the kth harmonic onwards, then, taking into account the relationship $2\pi(kn_p \pm w)/M = 2\pi k/(p \pm \Delta p)$, we see that the uncertainty is decreased, $\Delta p \sim 2wp^2/(kM)$.

The main source of random modifications of repeats are point mutations, which create a random background without changing the period. The random insertions and deletions result in variations of the period. In order to understand qualitatively the role of these effects, let us average the harmonic $\exp(iq_k p)$, where p is a random value, with respect to the Gaussian distribution function with mean value p_0 and standard deviation $\langle(\Delta p)^2\rangle$,

$$\langle\exp(iq_k p)\rangle \\ \approx \left[2\pi\langle(\Delta p)^2\rangle\right]^{-1/2}\int_{-\infty}^{\infty} dp\,\exp\left[-\frac{(p-p_0)^2}{2\langle(\Delta p)^2\rangle} + iq_k p\right] \\ = \exp\left[iq_k p_0 - \frac{q_k^2\langle(\Delta p)^2\rangle}{2}\right]. \tag{4.7}$$

From (4.7) we see that random variations of the period result in stronger damping of harmonics with larger wave numbers. As a result, in particular, only a few first harmonics or only the first harmonic in a series may appear to be statistically significant.

Thus, the hidden periodicity in a sequence can be revealed by analysis of separate high peaks, the sums of equidistant harmonics, or by combining both methods.

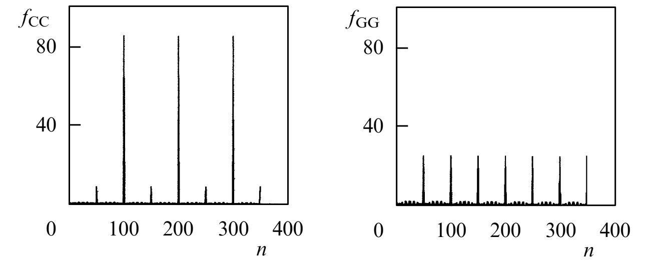

Figure 8. Half-spectra for the normalized harmonics for a sequence composed of 100 repeats ATAAACT after random substitutions of C by G in approximately a half of the even repetitions.

4.2 Criterion for statistical significance of a hidden periodicity

In the following it is assumed that only the harmonics $F_{\alpha\alpha}(q_n)$ are considered and the amplitudes of harmonics are normalized in accordance with (3.40). To obtain the criteria for statistical significance of hidden periodicities, it is necessary to use the grand multivariate probability density for a whole half-spectrum [see (3.11)],

$$p(f_1, \ldots, f_N) = \exp(-f_1 - \ldots - f_N)\,. \tag{4.8}$$

For a number of identical tests the extreme value statistics [100] should be taken into account, because the probability of finding the greater value is higher for a longer series of random numbers.

To begin with, consider the criteria for separate high peaks and then, following [21], we consecutively complicate the situation. The probability that the amplitudes of all harmonics do not exceed a threshold value f, i.e., $0 \leqslant f_1 \leqslant f, \ldots, 0 \leqslant f_N \leqslant f$, is given by

$$P(f_n \leqslant f;\, N) = \left[1 - \exp(-f)\right]^N, \tag{4.9}$$

and the probability that at least one of N harmonics exceeds f, is complementary to (4.9),

$$P(f_n > f;\, N) = 1 - \left[1 - \exp(-f)\right]^N. \tag{4.10}$$

In applications, the 10% or 5% levels of statistical significance are commonly used for $P(f_n > f; N)$. To obtain an approximate estimate, we can use the mean value and characteristic variance. Since, by definition, the moments $\langle f_{\max}^m \rangle$ are given by

$$\langle f_{\max}^m \rangle = \int_0^\infty \mathrm{d}f\, f^m\, \frac{\mathrm{d}P(f_n \leqslant f; N)}{\mathrm{d}f}\,, \tag{4.11}$$

we obtain

$$\langle f_{\max} \rangle = \sum_{k=1}^{N} \frac{1}{k}\,, \tag{4.12}$$

$$\left\langle (\Delta f_{\max})^2 \right\rangle = \langle f_{\max}^2 \rangle - \langle f_{\max} \rangle^2 = \sum_{k=1}^{N} \frac{1}{k^2}\,. \tag{4.13}$$

Equation (4.12) is a well-known result for ordered statistics for random values with the multivariate Rayleigh distribution [51]. If $N \gg 1$, then we have $\langle f_{\max} \rangle \approx \ln N$ in accordance with estimate (3.43), while $\left\langle (\Delta f_{\max})^2 \right\rangle$ approaches $\pi^2/6$.

The sum of r harmonics, S_r, has the probability density distribution [51]

$$p_r(S) = \int_0^\infty \mathrm{d}f_1 \ldots \int_0^\infty \mathrm{d}f_r\, \delta(S - f_1 - \ldots - f_r) \times \exp(-f_1 - \ldots - f_r) = \frac{S^{r-1}}{(r-1)!} \exp(-S)\,, \tag{4.14}$$

and the probability that S_r exceeds a threshold S is equal to

$$P_r(S) \equiv P(S_r > S) = \int_S^\infty \mathrm{d}S'\, p_r(S') = \exp(-S) \sum_{k=0}^{r-1} \frac{S^k}{k!}\,. \tag{4.15}$$

From (4.14) we obtain the mean value and variance,

$$\langle S_r \rangle = r\,, \quad \left\langle (\Delta S_r)^2 \right\rangle = r\,. \tag{4.16}$$

If we analyze N_r sums S_r such that each sum does not include any harmonic contained in the other sums, then in accordance with (4.9) and (4.10) we obtain that the probability of exceeding S by at least one of the sums is given by

$$P(S_r > S; N_r) = 1 - \left[1 - P_r(S)\right]^{N_r}. \tag{4.17}$$

The expressions for the mean value and variance for a sum of N_r different series are rather cumbersome and it is easier to use (4.17) directly.

In the technique with windows [see (4.6)], the probability that the sum of r harmonics does not exceed a threshold S is given by

$$P_w(S_r \leqslant S) = \int_0^S \mathrm{d}f_1\, p_w(f_1) \int_0^{S-f_1} \mathrm{d}f_2\, p_w(f_2) \ldots \int_0^{S-f_1-\ldots-f_{r-1}} \mathrm{d}f_r\, p_w(f_r)\,, \tag{4.18}$$

$$p_w(f) = \frac{\mathrm{d}P(f_n \leqslant f; 2w+1)}{\mathrm{d}f}\,,$$

$$P(f_n \leqslant f; 2w+1) = \left[1 - \exp(-f)\right]^{2w+1}. \tag{4.19}$$

Working with expressions like (4.18), it is convenient to use the Laplace transform,

$$P_w(S_r \leqslant S) = \int_{a-\mathrm{i}\infty}^{a+\mathrm{i}\infty} \frac{\mathrm{d}\lambda}{2\pi\mathrm{i}}\, \frac{\exp(\lambda S)}{\lambda} \left(\prod_{k=1}^{2w+1} \frac{k}{\lambda + k} \right)^r, \tag{4.20}$$

where $a > 0$. Thereby we get

$$\langle S_{r,w} \rangle = r \sum_{k=1}^{2w+1} \frac{1}{k}\,, \tag{4.21}$$

$$\left\langle (\Delta S_{r,w})^2 \right\rangle = r \sum_{k=1}^{2w+1} \frac{1}{k^2}\,, \tag{4.22}$$

in accordance with (4.12) and (4.13). Indeed, the mean value of the highest harmonic within the window of width w is given by (4.12) and the mean value of the sum of r highest harmonics is determined by (4.21). We can easily generalize (4.21) and (4.22) to consider the cases where different harmonics in an equidistant set are surrounded by windows of different widths.

Finally, we dwell upon the statistical significance for structural subharmonics. Consider two sums of r_1 and r_2 harmonics, respectively. Further, suppose that the sums do not contain any common harmonic and $S_{r_1} + S_{r_2} = S$ is fixed. Then the probability that the inequality $S_{r_1} \geqslant S_1$ holds under the above condition is given by [51]

$$P(S_{r_1} \geqslant S_1 | S_{r_1} + S_2 = S) = \frac{\int_{S_1}^{S} \mathrm{d}S_1'\, p_{r_1}(S_1') p_{r_2}(S - S_1')}{\int_0^S \mathrm{d}S_1'\, p_{r_1}(S_1') p_{r_2}(S - S_1')}\,, \tag{4.23}$$

where $p_{r_1}(S_1)$ and $p_{r_2}(S_2)$ are calculated in line with (4.14). The mean values $\langle S_1^m \rangle$ are determined in the usual manner,

$$\langle S_1^m \rangle = -\int_0^S \mathrm{d}S_1\, S_1^m\, \frac{\mathrm{d}P}{\mathrm{d}S_1}\,, \tag{4.24}$$

in particular,

$$\langle S_1 \rangle = \frac{r_1}{r_1 + r_2}\, S\,, \tag{4.25}$$

$$\left\langle (\Delta S_1)^2 \right\rangle = \frac{r_1 r_2}{(r_1 + r_2)^2 (r_1 + r_2 + 1)}\, S^2\,. \tag{4.26}$$

Consider now a set of equidistant harmonics $\{f_{n'k'}\}$, $k' = 1, \ldots, r$, with the sum S_r being statistically significant if the probability is calculated in accordance with (4.17). In this set, choose a subset of harmonics $k' = k, 2k, \ldots, r_1 k \leqslant r$ with the sum S_{r_1}. Then we can consider the set $\{f_{n'k'}\}$ as the kth subharmonic, if the inequality

$$\frac{S_{r_1}}{S_r} > \frac{r_1}{r} + 2\left[\frac{r_1(r - r_1)}{r^2(r+1)}\right]^{1/2}, \tag{4.27}$$

holds and $S_r - S_{r_1}$ is a statistically significant sum provided that the probabilities are calculated in accordance with (4.15) for $r - r_1$ harmonics. In general, these conditions are only necessary and imply that the distributions of harmonic amplitudes in both groups of r_1 and $r - r_1$ harmonics are approximately uniform (cf. Fig. 8).

4.3 Periodicity $p = 3$ is a fundamental property of protein-coding regions

The detailed analysis of all hidden periodicities in genomic DNA is rather cumbersome [21, 26]. As was noted above, the mechanism of period multiplication plays an important role in the formation of a periodicity. Another important factor is the effect of commensurability – incommensurability, when the incommensurable periods p_1 and p_2 generate the statistically significant periods $P = n_1 p_1 \approx n_2 p_2$, where n_1 and n_2 are integers. Usually these effects manifest themselves only approximately and in order to reveal them, we have to analyze all four spectra $f_{\alpha\alpha}(q_n)$, $\alpha \in (\mathrm{A, C, G, T})$. Thus, care should be taken when searching for hidden periodicities using the sum $\sum_\alpha f_{\alpha\alpha}(q_n)$ [82, 83, 85] or the product $\prod_\alpha f_{\alpha\alpha}(q_n)$ [101].

As an example, we consider the periodicity $p = 3$ which is a universal feature of protein-coding regions [5, 41, 45] (this property was checked by scanning all coding sequences in the database). Since for PHIX174 almost all the DNA sequence codes for proteins, the corresponding peaks at $n = 1795$ are clearly seen in Fig. 3. With a few exceptions discussed in the following, the peaks are retained for any of the three binary sequences, R – Y, W – S, K – M (Fig. 9). We take into account that in accordance with (3.13) for binary sequences the following equalities hold: $F_{\mathrm{RR}}(q_n) = F_{\mathrm{YY}}(q_n)$, $n \neq 0$, etc. The corresponding maximum relative heights (3.40) and the harmonic numbers for different binary sequences are $f_{\mathrm{RR,\,max}} = 33.8$, $n = 1795$; $f_{\mathrm{WW,\,max}} = 35.0$, $n = 1797$; $f_{\mathrm{MM,\,max}} = 35.7$, $n = 1797$. The harmonic with $n = 1795$ for the R – Y sequence is the closest to the strict periodicity $p = 3$.

Like any other pronounced periodicity, the periodicity $p = 3$ in the PHIX174 genome generates a cascade of structural subharmonics [21]. For periods divisible by 3, Fig. 10 shows the maximum ratios $(S_{r,w} - \langle S_{r,w}\rangle)/\langle(\Delta S_{r,w})^2\rangle^{1/2}$, which are calculated as follows. The high harmonics with numbers $1787 \leqslant n \leqslant 1803$ in the vicinity of the period $p = 3$ are excluded from the spectrum. The subharmonics with the periods $3k$ $(k = 2, \ldots, 20)$ correspond to the set of equidistant harmonics beginning from the number n giving the ratio M/n that is the nearest to $3k$, then, these harmonics are surrounded by windows of the same width w [see (4.6)]. The average values of the sums and the standard deviations for the sums of harmonics for random sequences are calculated in line with (4.21) and (4.22). Further, the widths of the windows are allowed to vary within the ranges $w = 0 – 10$ $(k = 2 – 9)$; $0 – 8$ $(k = 10)$; $0 – 5$ $(k = 11)$; $0 – 4$ $(k = 12)$; $0 – 3$ $(k = 13, 14)$; $0 – 2$ $(k = 15 – 20)$, then, the width is chosen such that $(S_{r,w} - \langle S_{r,w}\rangle)/\langle(\Delta S_{r,w})^2\rangle^{1/2}$ attains its maximum. Using the criteria from Section 4.2, we find that the subharmonics with periods $p = 6$, 18, and 54 are statistically significant. The peculiar role of the period $p = 6$ may likely be due to the contribution of the β-sheets in the protein structure (see Table 1).

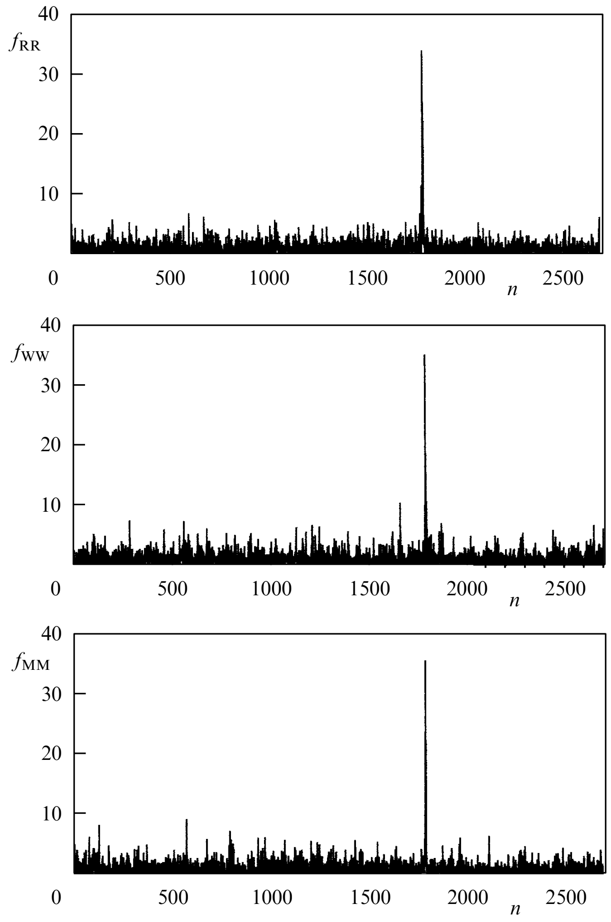

Figure 9. Half-spectra for the normalized harmonics for the binary subdivisions R – Y, W – S, and M – K in the genomic sequence of the bacteriophage PHIX174.

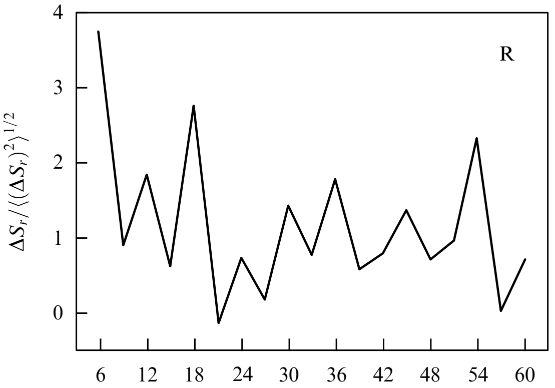

Figure 10. Normalized deviations for the sums of equidistant harmonics corresponding to subharmonics with periods divisible by 3 for the R – Y subdivision in the genome of PHIX174.

Several hypotheses were suggested to explain the origin and universality of the periodicity $p = 3$. The fundamental role of this periodicity may be due to the evolutionary origin of the genetic code and the domination of RRY [96] or RNY [97] codons (R is a purine, Y is a pyrimidine, and N is any nucleotide) at the early stage of molecular evolution. It is essential that the periodicity $p = 3$ is also observed in transport RNA (tRNA) [102], which plays an important role in the synthesis of proteins on the messenger RNA template [29]. In Ref. [103] the hypothesis was suggested that the codons GCT are of great importance in determining the correct framework for reading mRNA during protein synthesis, however, the hypothesis fails to be proved by experiments [104].

Due to the degeneracy of the genetic code and the fact that the third positions in codons are of minor importance, frequently these positions are occupied predominantly by a nucleotide of the same type [105–109, 57]. To illustrate this phenomenon, we choose an example for which this effect is especially pronounced. The collagen helix is composed of repeating subunits consisting of three amino acids [29, 33, 34]. The first position is usually occupied by a glycine, which is coded by the triplet GGN (G is a guanine), in the second position there is predominantly a proline coded by triplets CCN (C is a cytosine), and the third position is filled rather arbitrarily. For coding DNA this means that the periodicity $p = 9$ should be observed for nucleotides G and C. The corresponding half-spectrum for $f_{\mathrm{CC}}(q_n)$ (accession number for the DNA sequence is X52046 in the EMBL database) is shown in Fig. 11, where we observe a typical series of equidistant peaks. However, the spectrum $f_{\mathrm{TT}}(q_n)$ again has a very high peak corresponding to $p = 3$. The example shows that a thymine dominates in the third positions in codons GGN and CCN.

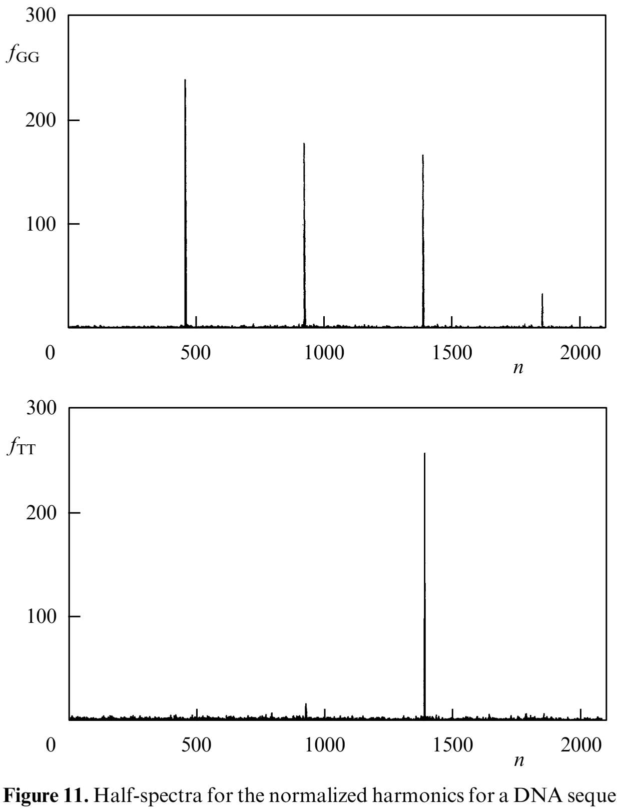

Figure 11. Half-spectra for the normalized harmonics for a DNA sequence coding for a fibrillar protein collagen.

For the most part, the DNA of PHIX174 codes for globular proteins, while Fig. 11 shows the spectra for a sequence coding for a fibrillar protein with a strongly elongated structure [33, 34]. For completeness, we also show a spectrum for a sequence coding for a transmembrane protein, in which the polypeptide chain pierces a membrane several times, so the part of the chain is within the intercellular space, the other part is in the cytoplasma, and the rest is within the membrane. The half-spectrum for the corresponding sequence (accession number is U59464) coding for the transmembrane protein PTCH is shown in Fig. 12. Only a half-spectrum for A is presented, because all the spectra are quite analogous. In the human genome there is a unique copy of this gene, and mutations in it result in basal-cell cancer [110, 111]. The data on the distribution of hydrophobicity along the protein chain and the location of specific indicator sequences lead to the conclusion that there are six cytoplasmatic domains, five extracellular domains, and ten transmembrane domains in the protein molecule [112]. Thus, in addition to a characteristic peak for the period $p = 3$, in the region of small wave numbers we observe several high harmonics which are due to the segmentation and mosaic structure of the protein considered (see Fig. 12).

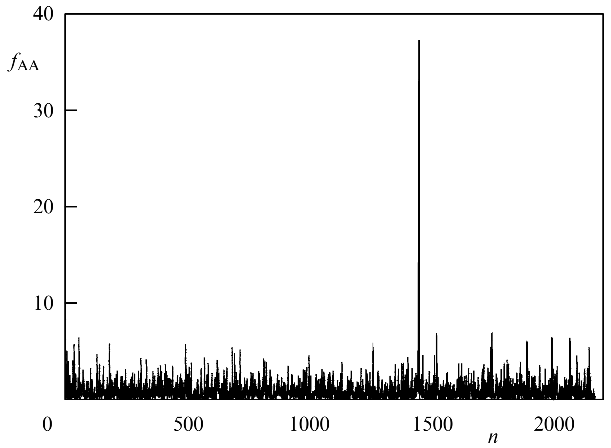

Figure 12. Half-spectra for the normalized harmonics $f_{\mathrm{AA}}(q_n)$ for a DNA sequence coding for a transmembrane protein PTCH.

One more mechanism for generation of the periodicity $p = 3$ was suggested in Refs [18, 21, 27]. Consider a double-stranded DNA where a strand composed of purines R is positioned opposite the complementary strand composed of pyrimidines Y. Suppose that pyrimidines penetrate between purines analogously to a zipper-like mechanism,

$$\begin{matrix}\mathrm{R}\ldots\mathrm{R}\\ \mathrm{Y}\ldots\mathrm{Y}\end{matrix} \rightarrow \mathrm{RY}\ldots\mathrm{RY}\,.$$

Suppose further that the process is consecutively repeated,

$$\begin{matrix}\ldots\mathrm{R}\ldots\mathrm{Y}\ldots\\ \ldots\mathrm{Y}\ldots\mathrm{R}\ldots\end{matrix} \rightarrow \ldots\mathrm{RY}\ldots\mathrm{YR}\ldots,$$

etc. It is easily seen that these iterations can be represented as the consecutive substitutions R → RY, Y → YR. As a result, we obtain one of the well-known substitutional sequences, namely, the so-called Thue–Morse sequence, which was introduced for the first time in number theory in 1906. The spectral characteristics for this sequence are well investigated

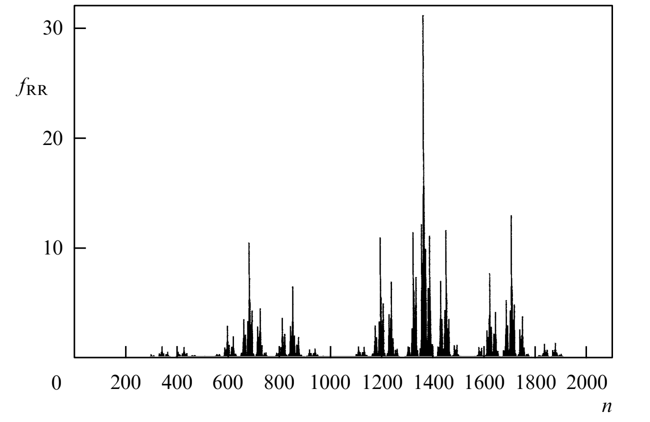

Figure 13. Half-spectra for the normalized harmonics for a Thue – Morse sequence obtained after 12 iterations R → RY, Y → YR beginning with R.

[113, 114]. In particular, it has been proved that the periodicity $p = 3$ dominates. In such a model the mechanism of period multiplication is realized. Fig. 13 shows the spectrum for a sequence obtained after 12 consecutive iterations beginning from R ($M = 4096$). We should also point out several characteristic peaks in the region between the periods $p = 2$ and $p = 3$, which are often observed in DNA sequences. In the probabilistic realization of the model [115], the spectra in the region of small wave numbers are analogous to that of $1/q^{\nu}$-noise, thereby simulating the effects of segmentation.

Thus, within the framework of the given mechanism, the periodicity $p = 3$, the period multiplication, and other features are the direct consequence of the complementarity.

The fragments from triplet repeats also play important structural and regulatory roles in the human genome. The proliferation of repeats (the increase in fragment lengths) can result in serious genetic diseases [116 – 118]. However, it is not clear yet how specific in this sense (as against their function in protein-coding regions) is the role of fragments consisting of triplet repeats as compared with the numerous other families of repeats.

4.4 Other periodicities

Of all other hidden periodicities in the genomic DNA sequences, the periodicity related to the pitch of a DNA double helix appears to be the most typical [38, 85, 119]. For the exon-intron sequences in the human and mouse genomes it was shown that in the histograms for distances between the 5′ end of an intron and the 3′ ends of the first and all subsequent exons there are quasiperiodic variations of a period $p \approx 204-205$ [120]. The analysis of hidden periodicities in exon-intron sequences prove that this periodicity is quite frequently reproduced [26]. However, to be more precise, we must speak about a set of approximately reproducible periodicities within this range of lengths rather than a unique periodicity. These periodicities are probably related to the packing of DNA in nucleosomes. Since there exist a variety of ways of molecular evolution, in different genomes we observe that the periodicity $p \approx 204-205$ manifests itself in different ways for different types of nucleotides and can also be realized by the mechanism of period multiplication. Also noteworthy is the role of the commensurability – incommensurability effects, indeed, the periodicity $p \approx 204-205$ covers approximately 20 pitches of a DNA double helix in the B form.

In Ref. [121] it was pointed out that the hidden periodicity $p = 2$ occurs frequently in introns. The fact indicates on the modifications of the Z form and the possibility for dynamic B – Z transitions to occur (see Table 1). Fragments composed of purines and pyrimidines alternating in a regular manner are also found only in introns [122]. For flanking 5′ and 3′ fragments in exon-intron regions both periodicities $p = 2$ and $p = 3$ are frequently observed [123]. However, these hidden periodicities can be revealed only statistically and are not well reproducible as compared to the periodicity $p = 3$ in protein-coding regions. It is interesting to note that the periodicity $p = 2$ occurs at the first iteration for the Thue – Morse sequence.

Let us consider once again the genome of bacteriophage PHIX174. The inserts in Fig. 3 and Eqn (3.48) indicate that there exist five large superperiods in the sequence considered. Figure 14 shows the ratios $(S_{r,w} - \langle S_{r,w}\rangle)/\langle(\Delta S_{r,w})^2\rangle^{1/2}$ for equidistant series, which are consecutively generated beginning from $n = 3-51$. The high harmonics within the range $1787 \leqslant n \leqslant 1803$ are removed from the spectrum, and all equidistant harmonics beginning from the third are surrounded by windows of width $w = 1$. Only the results for C are presented in Fig. 14 (for more details, see Ref. [21]). The peaks for almost all n divisible by 5 are clearly seen in Fig. 14 thereby confirming a conclusion about the existence of five large superperiods in the sequence. It is well-known that the protein envelope (the capsid), which surrounds the circular DNA of bacteriophage PHIX174, has the form of a regular dodecahedron [95]. It is not unlikely that the superperiods $p = M/5$ are related to the DNA packing and 5th-order symmetry axes for a dodecahedron [124].

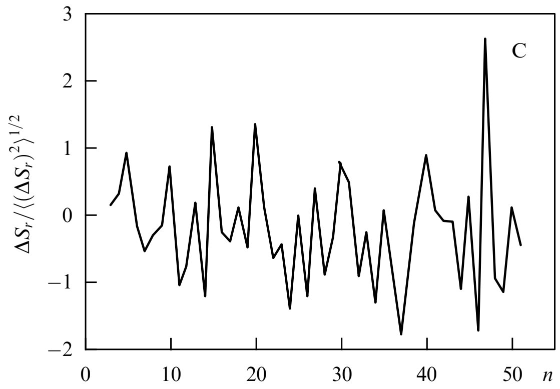

Figure 14. Normalized deviations for the sums of equidistant harmonics $f_{CC}(q_n)$ generated starting from a number n for the PHIX174 genome.

These examples illustrate the fact that the hidden periodicities in DNA sequences contain a great deal of evolutionary and structural information. Methods for reconstruction of the randomly modified hidden repeats in an explicit form are described in Refs [22, 25].

5. Analysis of correlations in genomic DNA sequences

At the first stage of analysis of DNA sequences the problem consists in splitting a sequence into segments with different functions (protein-coding regions, tRNA-coding regions, satellite DNA, etc.). Then, at the second stage there is a problem concerning the structural coupling between the different segments. In this section we consider the application of spectral methods to the analysis of correlations between nucleotide positions in a sequence.

5.1 Method of sweeps in the spectra for structure factors and correlation functions

To begin with, consider a simple case without preliminary separation of the segmentation effects. The large-scale density variations on scales comparable with the total length of a sequence can be investigated with the help of the smoothed standard deviations for correlation functions [see (3.53) and Fig. 6] or the smoothed Fourier spectra:

$$\widetilde{f}_{\alpha\alpha}(q_n) = (2k+1)^{-1} \sum_{n'=n-k}^{n+k} f_{\alpha\alpha}(q_{n'}) \,, \tag{5.1}$$

where $f_{\alpha\alpha}(q_{n'})$ are defined by (3.40). The statistical criteria for $\widetilde{f}_{\alpha\alpha}(q_n)$ can be found in Refs [17, 18]. Here we consider the problem in a more qualitative sense.

We explain the basic ideas with our standard example of the PHIX174 genome. Fig. 15 shows the smoothed half-spectra obtained for PHIX174 with $k = 200$ in (5.1). In the region of small wave numbers in the spectra for A and T we observe a steady tendency to increase as q_n decreases and in spectra for C and G a tendency (not quite regular) to decrease. Following the standard terminology [125], we call such behavior persistent in the first case and antipersistent in the second case, respectively. Suppose a window of width $W \sim 1/q_n$ is moving along the sequence, then, in the first case we observe a tendency for the character of variations to be preserved, while in the second case a tendency to change.

One of standard methods for the analysis of variations is the method of Hurst's curves (or the method of normalized sweep) [125]. Now we illustrate how the behavior of the smoothed spectra is related to the behavior of Hurst's curves [19]. Choose a site m and a window width m_0. Defining the average nucleotide density within the interval from m to $m + m_0 - 1$ as

$$\bar{\rho}_\alpha = m_0^{-1} \sum_{m'=m}^{m+m_0-1} \rho_{m',\alpha} \,, \tag{5.2}$$

and then introducing a deviation,

$$\Delta_\alpha(m, \tilde{m}) = \sum_{m'=m}^{\tilde{m}} (\rho_{m',\alpha} - \bar{\rho}_\alpha) \,, \qquad m \leqslant \tilde{m} \leqslant m + m_0 - 1, \tag{5.3}$$

we determine a sweep

$$R_\alpha(m, m+m_0-1) = \max_{m \leqslant \tilde{m} \leqslant m+m_0-1} \Delta_\alpha(m, \tilde{m}) - \min_{m \leqslant \tilde{m} \leqslant m+m_0-1} \Delta_\alpha(m, \tilde{m}). \tag{5.4}$$

Then, the sweep $R_\alpha(m, m+m_0-1)$ is divided by the standard deviation,

$$\sigma(\rho_\alpha) = \left[m_0^{-1} \sum_{m'=m}^{m+m_0-1} (\rho_{m',\alpha} - \bar{\rho}_\alpha)^2 \right]^{1/2} , \tag{5.5}$$

and the ratio $R_\alpha/\sigma(\rho_\alpha)$ for each current value of m_0 is averaged over all sites m. The average ratio $\overline{R_\alpha/\sigma(\rho_\alpha)}$ is compared with the analogous value $\langle \overline{R_\alpha/\sigma(\rho_\alpha)} \rangle$ obtained by averaging over the ensemble of random sequences with the same nucleotide composition.

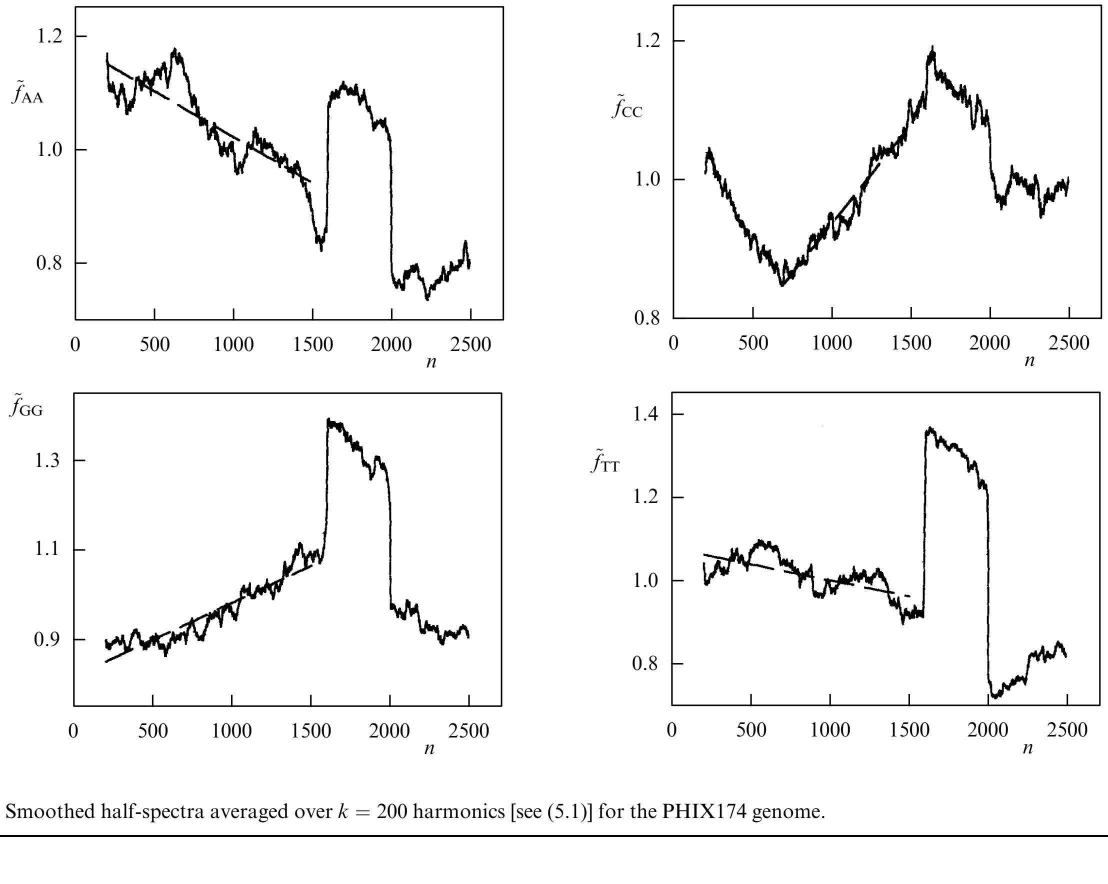

Figure 15. Smoothed half-spectra averaged over $k = 200$ harmonics [see (5.1)] for the PHIX174 genome.

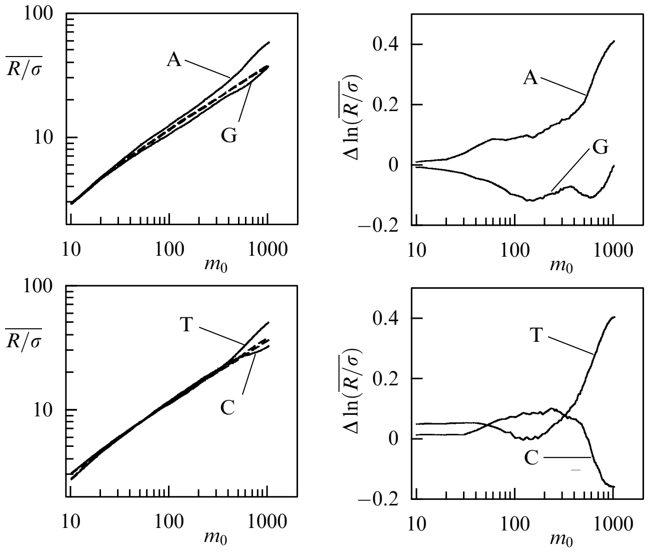

Figure 16. Hurst's curves for the PHIX174 genome. Also shown are Hurst's curves obtained by averaging over 20 random realizations for sequences with the same nucleotide composition (dashed curves).

Figure 16 shows the dependences of $\overline{R_\alpha/\sigma(\rho_\alpha)}$, $\langle R_\alpha/\sigma(\rho_\alpha)\rangle$, and

$$\Delta\ln\left[\overline{R_\alpha/\sigma(\rho_\alpha)}\right] = \ln\left[\overline{R_\alpha/\sigma(\rho_\alpha)}\right] - \ln\left[\langle\overline{R_\alpha/\sigma(\rho_\alpha)}\rangle\right]$$

obtained for PHIX174 with $10 \leqslant m_0 \leqslant 1000$. We observe that the deviations $\Delta\ln\left[\overline{R_\alpha/\sigma(\rho_\alpha)}\right]$ for A and T are positive, while for G and C they are mostly negative in accordance with the behavior of the smoothed spectra in Fig. 15. Such behavior can be understood by considering a spectrum of type $F(q_n) \propto 1/q^{\nu}$ with a constant exponent ν. Then, for the normalized deviation we obtain $\overline{R/\sigma} \propto m_0^{(1+\nu)/2}$, hence, the sign of the exponent ν determines the deviations from the dependence for random sequences. The dashed curves in Fig. 16 are obtained by averaging over 20 random realizations. When averaging over m, we take into account that the PHIX174 genome is circular and the position functions $\rho_{m,\alpha}$ are replaced by the cyclically continued functions $\rho^{\mathrm{c}}_{m,\alpha}$ [see (3.8)]. The fluctuations in $\ln\left[\overline{R_\alpha/\sigma(\rho_\alpha)}\right]$ for random sequences increase with m_0 and are about $\left\langle\left\{\Delta\ln\left[\overline{R_\alpha/\sigma(\rho_\alpha)}\right]\right\}^2\right\rangle^{1/2} \approx 0.01$ for $m_0 = 100$ and 0.05 for $m_0 = 1000$, hence, we attain the conclusion that the observed deviations for $\Delta\ln\left[\overline{R_\alpha/\sigma(\rho_\alpha)}\right]$ are statistically significant. The antipersistent behavior of Hurst's curve for C is in a good agreement with the conclusion that there are about five large superperiods for C (see Fig. 14), because quasiperiodic variations may be one of the reasons for antipersistent behavior. The fact that the dependence of $\Delta\ln\left[\overline{R_\alpha/\sigma(\rho_\alpha)}\right]$ on m_0 is not monotone implies the existence of a number of characteristic lengths (in the case considered these lengths are about the gene sizes). Such non-monotonic behavior is typical for other genomic DNA sequences as well [15, 19, 63–65]. The curves in Fig. 16 show once again that we should begin the study with the analysis of all four nucleotides. Indeed, for binary subdivision R = (A, G) and Y = (C, T) the persistent and antipersistent behavior of variations for A and G, and, correspondingly, for T and C, can compensate each other thereby weakening the large-scale variations.

Now we proceed to the analysis of mutual correlations between nucleotides of different types. An integral estimate of the significance for correlations between various nucleotides can be obtained from (3.32), (3.36), and (3.38). For random sequences the probability that a deviation

$$x_{\alpha\beta} = \frac{k(F_{\alpha\alpha}|F_{\beta\beta};\ M-1) - \langle k_{\alpha\beta}\rangle}{\left[2\left\langle(\Delta k_{\alpha\beta})^2\right\rangle\right]^{1/2}} \tag{5.6}$$

is within the range $-x \leqslant x_{\alpha\beta} \leqslant x$ (or $|x_{\alpha\beta}| \leqslant x$), is given by

$$P\left(|x_{\alpha\beta}| \leqslant x\right) = 2\pi^{-1/2}\int_0^x \mathrm{d}x'\ \exp(-x'^2) \equiv \operatorname{erf}(x)\,, \tag{5.7}$$

then, the probability that $|x_{\alpha\beta}| > x$ is equal to

$$P\left(|x_{\alpha\beta}| > x\right) = 1 - \operatorname{erf}(x)\,. \tag{5.8}$$

Choosing the thresholds $P\left(|x_{\alpha\beta}| > x\right) = 0.1$ and 0.05, we find $x = 1.16$ and 1.39, respectively, and now we can assess the statistical significance of correlations observed in genomic DNA sequences.

In applications it is essential not only to reveal significant correlations or their absence but it is also important to identify the source of the significant correlations. From the structural point of view, the correlations may be caused by the coincident hidden periodicities, large-scale density variations, short-range coupling, or coherent point mutations. The origin of correlations can be revealed with the help of the method of sweeps in spectra for structure factors and correlation functions [23, 27]. In this method the correlation coefficients are calculated for a part of a spectrum, $k(F_{\alpha\alpha}|F_{\beta\beta};\ 1 \leqslant n' \leqslant n)$ and $k(K_{\alpha\alpha}|K_{\beta\beta};\ 1 \leqslant m_0' \leqslant m_0)$, where n and m_0 are consecutively increasing from 1 up to N [see (3.11)] (sweep from left to right). Analogously, the coefficients $k(F_{\alpha\alpha}|F_{\beta\beta};\ n \leqslant n' \leqslant N)$ and $k(K_{\alpha\alpha}|K_{\beta\beta};\ m_0 \leqslant m_0' \leqslant N)$ are calculated, where n and m_0 are consecutively decreasing from N to 1 (sweeps from right to left). We recall that for integral correlation coefficients the equality (3.35) holds, however, the analogous equality fails to be true for a part of a spectrum. Comparing the dependences for current correlation coefficients on n and m_0 when sweeping from left to right and in the opposite direction, we can find the contributions of different regions in the spectra and to reveal the main sources of correlations. For example, if there are jumps in dependences $k(F_{\alpha\alpha}|F_{\beta\beta};\ 1 \leqslant n' \leqslant n)$ and $k(F_{\alpha\alpha}|F_{\beta\beta};\ n \leqslant n' \leqslant N)$, we get the conclusion that there exist coincident (in the case of correlations) or shifted (in the case of anticorrelations) periodicities. Due to mutual complementarity between wave numbers q_n and scales m_0 for Fourier transform [see (3.6)], $m_0 \propto 1/q_n$, it is possible to study the large-scale correlation in detail using $k(K_{\alpha\alpha}|K_{\beta\beta};\ m_0 \leqslant m_0' \leqslant N)$, because such correlation coefficients contain only the correlation functions $K(m_0')$ with $m_0' \geqslant m_0$, while for studying short-scale correlations it is convenient to use $k(F_{\alpha\alpha}|F_{\beta\beta};\ n \leqslant n' \leqslant N)$, because these coefficients contain only the contributions of structure factors with wave numbers $q_{n'} \geqslant q_n \propto 1/m_0$.

Figure 17 (left panel) shows the dependences of $k_{\mathrm{rel}} = (k_{\alpha\beta} - \langle k_{\alpha\beta}\rangle)/\langle k_{\alpha\beta}\rangle$ [see (3.37)] in the case of correlations A–T in the PHIX174 genome. The 'sweeps' for $k_{\mathrm{rel}}(F_{\mathrm{AA}}|F_{\mathrm{TT}})$ reveal the important role of the coincident periodicity with $p = 3$ (cf. Fig. 3 and jumps at $n \approx 1795$ in Fig. 17). The dependence of $k_{\mathrm{rel}}(K_{\mathrm{AA}}|K_{\mathrm{TT}};\ m_0 \leqslant m_0' \leqslant N)$ on m_0 indicates the significance of large-scale correlations and a certain structural integrity of the genome as a whole. Indeed, if we split the genome into fragments of length

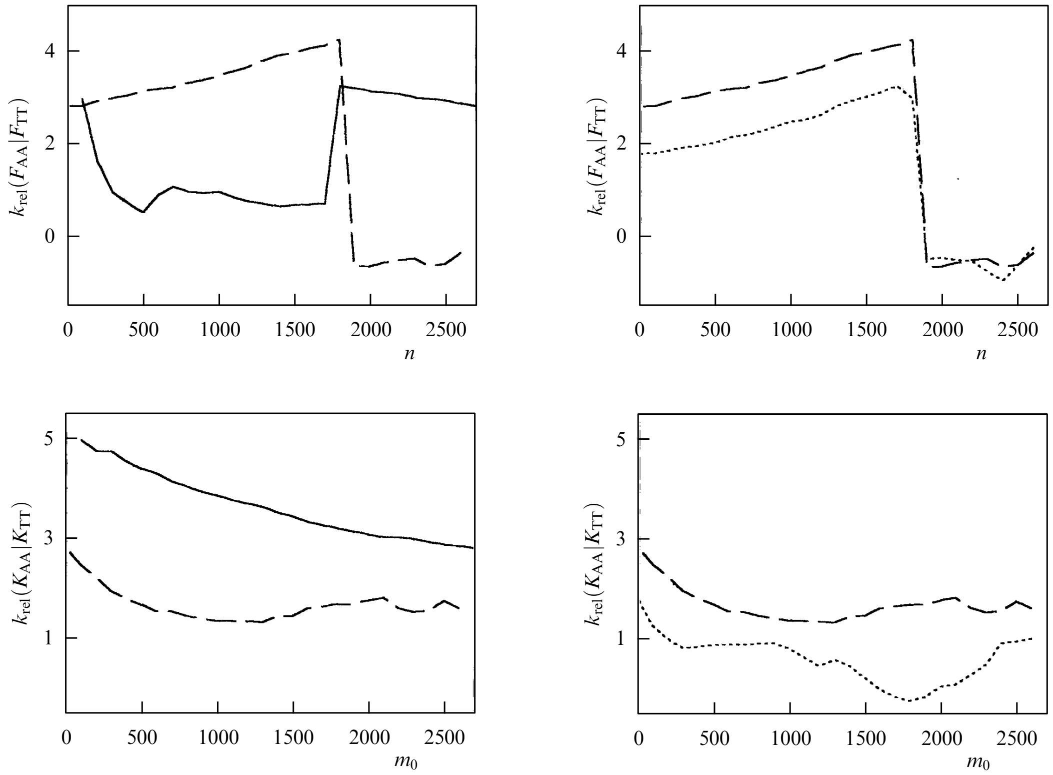

Figure 17. Current normalized correlation coefficients for the method of sweeps in the structural spectra for A – T correlations in the PHIX174 genome. The solid curves describe the inflation from left to right (left panels), while the dashed curves describe the inflation from right to left. The dependences describing the inflation from right to left for the PHIX174 genome after random reshuffling of the fragments of length $l = 500$ are shown by dotted curves together with the corresponding dependences obtained without reshuffling (right panels).

$l = 500$ (the mean gene size for PHIX174) and reshuffle the fragments in a random manner, we obtain the dependences shown in Fig. 17 (right panel). Additional growth for $k_{\text{rel}}(K_{\text{AA}}|K_{\text{TT}};\ m_0 \leqslant m_0' \leqslant N)$ at $m_0 \leqslant 500$ indicates the hierarchical structure of correlations, i.e., the correlations within the genes appear to be stronger than the correlations between the genes. In this sense a gene can be spoken about as a structural unit.

The dependences in Fig. 17 show that the periodicity $p = 3$ coherently runs through the clusters of several genes. To confirm this conclusion, there exists a more direct method consisting in studying the size-dependences for the three-periodic harmonics $f_{\alpha\alpha}(q_n)$ [see (3.40)] calculated for the subsequences of length L [20]. If structural coupling is absent, we observe a quasirandom variation in the vicinity of a constant value. On the contrary, if there exists coherent coupling, then the height of $f_{\alpha\alpha}(q_n)$ increases with increasing L. If the periodicity $p = 3$ is perfect, it follows that $f \propto L$. If there are two segments of lengths L_1 and L_2, respectively, and the periodicities $p = 3$ in the segments are shifted with respect to each other, then $f \propto (L_1^2 + L_2^2 - L_1 L_2)/(L_1 + L_2)$. In this case we observe a minimum at $L_2 = L_1(\sqrt{3} - 1)$ and a characteristic tooth-like dependence. Finally, if a segment of length L_{c} with the ideal periodicity $p = 3$ is adjoined to a segment of length L_r with a random nucleotide distribution, then $f \propto (L_{\text{c}}^2 + L_r)/(L_{\text{c}} + L_r)$.

These dependences for the binary sequences R – Y obtained for the genomes of the bacteriophages PHIX174 and MIG4XX (accession number J02454) and the RNA virus TOEAV (accession number X53459) are shown in Fig. 18. These curves indicate that there exists a structural coupling and block association between several genes (for a general discussion of similar mechanisms, see Refs [4, 5, 28, 29]). This effect is partially due to overlapping of some genes in the

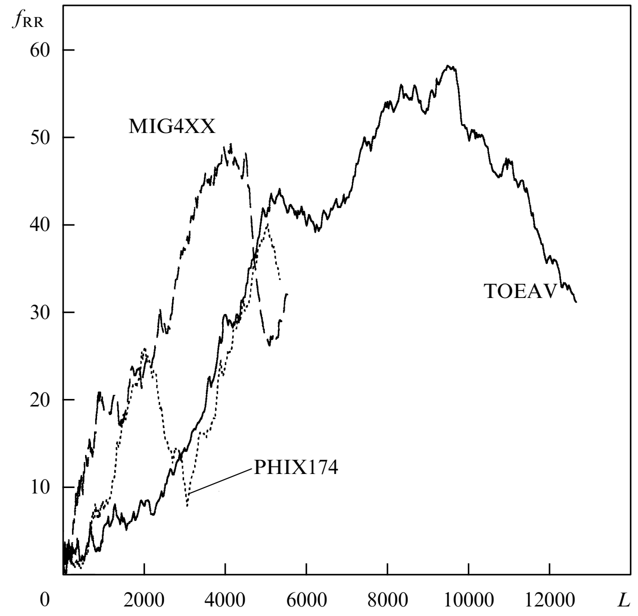

Figure 18. Size-dependence for the amplitudes of the normalized three-periodic harmonics $f_{\text{RR}}(q_n)$ calculated for successively increasing parts, $1 \leqslant m \leqslant L$, of the genomic sequences for PHIX174, MIG4XX, and TOEAV.

genomes considered. However, the blocks comprise non-overlapping genes as well. If the PHIX174 genome is split into segments of length $l = 500$ or 100 and the segments are then randomly reshuffled, we observe that the structural coupling is destroyed (Fig. 19).

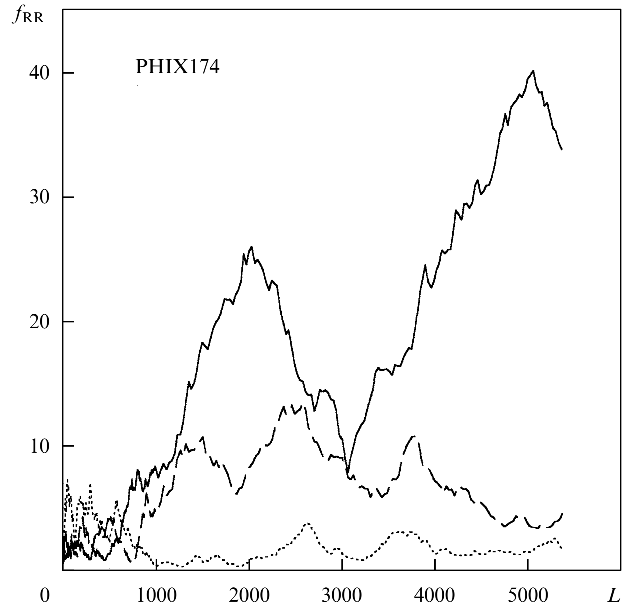

Figure 19. Size-dependence for the amplitudes of the normalized three-periodic harmonics $f_{\rm RR}(q_n)$ calculated for successively increasing parts, $1 \leqslant m \leqslant N$, of the genomic sequences for PHIX174 (solid curve) and for sequences obtained from the original genomic sequence for PHIX174 by splitting it into fragments of length $l = 500$ and 100 and subsequently reshuffling the fragments (dashed and dotted curves, respectively).

One reason why we chose the MIG4XX genome is that PHIX174 and MIG4XX occupy close positions on the evolutionary tree. Both bacteriophages attack a bacterium *Escherichia coli* and there is also a one-to-one correspondence between the genes of PHIX174 and MIG4XX. Nevertheless, the genomes differ strongly in their structural characteristics [19] and the nucleotide compositions of the corresponding DNA sequences. In accordance with the general concept [95], the viruses (bacteriophages represent a separate subgroup of viruses) can be considered as 'escaped' DNA. It is represented by short DNA fragments, which can exist rather independently, although for replication they require the proteins of a host cell. Additional analysis reveals that the large DNA fragments in PHIX174 and MIG4XX are mutually complementary [22, 25]. Since this evolutionary mechanism is quite universal [4, 41], it is useful to have it in mind when studying the correspondence between DNA sequences for evolutionary close species because the closeness of sequences is not necessarily straightforward.

5.2 Analysis of correlations in segmented DNA sequences

In the analysis of DNA sequences, the problem concerned with the structural coupling or the absence of such a coupling between different elements of the mosaic structure of a genome is of great importance. In this section we describe the methods for analysis of correlations, with the effects of segmentation being taken into account from the very beginning [27].

Let a DNA sequence of total length M be composed of K non-overlapping segments of lengths $\{L_k\}$, $(k = 1, \ldots, K)$,

$$M = \sum_{k=1}^{K} L_k \,. \tag{5.9}$$

Determine the mean nucleotide density within each segment,

$$\bar{\rho}_\alpha^{(k)} = \frac{N_\alpha^{(k)}}{L_k} \,, \tag{5.10}$$

and introduce the differential position functions [cf. (3.1)],

$$\tilde{\rho}_{m,\alpha} = \rho_{m,\alpha} - \bar{\rho}_{m,\alpha} \,, \tag{5.11}$$

where $\bar{\rho}_{m,\alpha} = \bar{\rho}_\alpha^{(k)}$ if the site m is within the kth segment. The Fourier harmonics are defined as follows:

$$\rho_\alpha(q_n) = M^{-1/2} \sum_{m=1}^{M} \tilde{\rho}_{m,\alpha} \exp(-\mathrm{i} q_n m) \,,$$

$$q_n = \frac{2\pi n}{M} \,, \qquad n = 0, 1, \ldots, M-1 \,, \tag{5.12}$$

and all the other relationships are analogous to (3.5), (3.6), (3.10), and (3.11).

The mean values for diagonal elements of the structure factor $\langle F_{\alpha\alpha}(q_n) \rangle$ and correlation functions $\langle K_{\alpha\alpha}(m_0) \rangle$ are obtained by averaging over the ensemble of random sequences with the same lengths of segments $\{L_k\}$ and the same nucleotide composition within each segment as in the genomic sequence under consideration. The averaging within each segment is performed independently and is defined by (3.27) and (3.28). As a result, we obtain

$$\langle F_{\alpha\alpha}(q_n) \rangle = M^{-1} \sum_{k=1}^{K} L_k \bar{F}_{\alpha\alpha}^{(k)} \left[1 - \frac{\sin^2(q_n L_k/2)}{L_k^2 \sin^2(q_n/2)} \right] , \tag{5.13}$$

$$\bar{F}_{\alpha\alpha}^{(k)} = \frac{N_\alpha^{(k)} (L_k - N_\alpha^{(k)})}{L_k (L_k - 1)} \tag{5.14}$$

and

$$\langle K_{\alpha\alpha}(m_0) \rangle = -\sum_{k=1}^{K} \frac{\widetilde{K}(m_0, L_k) \bar{F}_{\alpha\alpha}^{(k)}}{L_k} \,, \tag{5.15}$$

$$\widetilde{K}(m_0, L) = \theta(L - m_0) \, \frac{L - m_0}{M} + \theta(L + m_0 - M) \, \frac{L + m_0 - M}{M} \,, \tag{5.16}$$

where $\theta(x)$ is the Heaviside step function, $\theta(x) = 1$ if $x > 0$ and $\theta(x) = 0$ otherwise. Then, for the deviations

$$\Delta F_{\alpha\alpha}(q_n) = F_{\alpha\alpha}(q_n) - \langle F_{\alpha\alpha}(q_n) \rangle \,,$$

$$\Delta K_{\alpha\alpha}(m_0) = K_{\alpha\alpha}(m_0) - \langle K_{\alpha\alpha}(m_0) \rangle \tag{5.17}$$

we calculate the correlation coefficients for a part of a spectrum and use the method of sweeps as was described in Section 5.1. For the random segmented sequences the mean values for integral correlation coefficients are given by

$$\langle k(F_{\alpha\alpha}|F_{\beta\beta};\ 1\leqslant n'\leqslant N)\rangle \equiv \langle k_{\alpha\beta}\rangle = \frac{\bar{F}^2_{\alpha\beta}}{\bar{F}_{\alpha\alpha}\bar{F}_{\beta\beta}}\,, \tag{5.18}$$

$$\bar{F} = M^{-1}\sum_{k=1}^{K}\bar{F}^{(k)}L_k\,, \tag{5.19}$$

$$\langle(\Delta k_{\alpha\beta})^2\rangle \approx \frac{1}{N}\,, \tag{5.20}$$

where the mean value $\bar{F}^{(k)}$ for each segment is calculated in line with (3.14), where M and N_α are replaced by L_k and $N_\alpha^{(k)}$, respectively [cf. also (5.14)].

The method is easily generalized for cases when segments belong to different classes. For definiteness, we consider alternating exon-intron fragments [29, 31]. An initial sequence will be denoted as eiei... Then, we define the differential position functions of two types, $\tilde{\rho}^{(e)}_{m,\alpha}$ and $\tilde{\rho}^{(i)}_{m,\alpha}$, where $\tilde{\rho}^{(e)}_{m,\alpha}$ is given by (5.11) if the mth site is within an exon and vanishes otherwise. Such a function $\tilde{\rho}^{(e)}_{m,\alpha}$ defines a sequence e0e0... with intron fragments replaced by zeroes. Similarly, we define a function $\tilde{\rho}^{(i)}_{m,\alpha}$ and a sequence 0i0i... Then, using the functions $\tilde{\rho}^{(e)}_{m,\alpha}$ and $\tilde{\rho}^{(i)}_{m,\alpha}$, the Fourier harmonics $\tilde{\rho}^{(e)}_\alpha(q_n)$ and $\tilde{\rho}^{(i)}_\alpha(q_n)$ are calculated [cf. (5.12)], and so on. To obtain the mean values $\langle F^{(e)}_\alpha(q_n)\rangle$ and $\langle K^{(e)}_\alpha(m_0)\rangle$, the summation in (5.13) and (5.15) is now performed only over exon segments, and analogously, when calculating $\langle F^{(i)}_\alpha(q_n)\rangle$ and $\langle K^{(i)}_\alpha(m_0)\rangle$, only intron segments are taken into account. Splitting a sequence into e0e0... and 0i0i... allows one to study the correlations not only between the separate segments but also to reveal the position effects related to the distribution of lengths $\{L_k\}$.

The correlation coefficients

$$k\big(F^{(e)}_{\alpha\alpha}|F^{(i)}_{\beta\beta};\ 1\leqslant n\leqslant N\big) = k\big(K^{(e)}_{\alpha\alpha}|K^{(i)}_{\beta\beta};\ 1\leqslant m_0\leqslant N\big)$$

and their analogues in the method of sweeps are of primary interest. For the random segmented sequences we have the following estimates:

$$\langle k(F^{(e)}_{\alpha\alpha}|F^{(i)}_{\beta\beta};\ 1\leqslant n\leqslant N)\rangle \equiv \langle k^{(e-i)}_{\alpha\beta}\rangle = 0\,, \tag{5.21}$$

$$\langle(\Delta k^{(e-i)}_{\alpha\beta})^2\rangle \approx \min\left(\frac{2}{L_e}\,,\ \frac{2}{L_i}\right), \tag{5.22}$$

$$L_e = \sum_{\text{exons}} L^{(e)}_k\,, \qquad L_i = \sum_{\text{introns}} L^{(i)}_k\,. \tag{5.23}$$

In this case the total number of random degrees of freedom is approximately equal to $\max(L_e, L_i)$ as is seen from the expression (5.22) for the variance. The statistical significance for mutual correlations is estimated in terms of the Gaussian variables

$$x^{e-i}_{\alpha\beta} = \frac{k\big(F^{(e)}_{\alpha\alpha}|F^{(i)}_{\beta\beta};\ 1\leqslant n\leqslant N\big)}{\big[2\langle(\Delta k^{(e-i)}_{\alpha\beta})^2\rangle\big]^{1/2}}\,. \tag{5.24}$$

The probability that the inequality $|x_{\alpha\beta}| > x$ holds for at least one of s independent Gaussian variables is given by

$$P\big(|x_{\alpha\beta}| > x;\ s\big) = 1 - \big[\operatorname{erf}(x)\big]^s\,. \tag{5.25}$$

For $s = 16$ we obtain from this equation that the 10% and 5% levels of statistical significance correspond to the thresholds $x = 1.92$ and 2.04, respectively.

As an example, we consider the correlations between the sequences e0e0... and 0i0i... for the collagen gene (accession number X52046) without the 5′ and 3′ flanking exon–intron pairs. Note that in this notation the spectra in Fig. 11 correspond to the sequence ee... obtained by splicing exon fragments. The sequence under consideration is composed of 49 exons (with lengths $\{L^{(e)}_k\}$ ranging from 54 to 295) and 48 introns (with lengths $\{L^{(i)}_k\}$ from 82 to 1561), $L_e = 4163$, $L_i = 22448$, $M = 26611$ [see (5.23)]. The integral correlation coefficients are given in Table 2. The values of normalized variables (5.24) are given in parentheses.

Table 2. Cross correlation coefficients for structural coupling between exons and introns in the collagen gene (X52046).

	Ae	Ge	Te	Ce
Ai	−0.002	−0.003	−0.018	0.006
	(0.12)	(0.23)	(1.36)	(0.46)
Gi	0.018	0.014	0.008	−0.002
	(1.33)	(1.05)	(0.60)	(0.17)
Ti	0.015	0.006	−0.001	0.018
	(1.15)	(0.48)	(0.08)	(1.37)
Ci	0.034	0.022	0.045	0.021
	(2.57)	(1.68)	(3.38)	(1.57)

Corresponding to the strongest correlations Te–Ci, the dependences obtained by the method of sweeps are shown in Fig. 20. The dependences for $k(F^{(e)}_{\mathrm{TT}}|F^{(i)}_{\mathrm{CC}}) \equiv k(F;\ \text{Te–Ci})$ demonstrate once again the structural coupling through the periodicity $p = 3$ (see jumps at $n \approx 8870$). Analyzing the

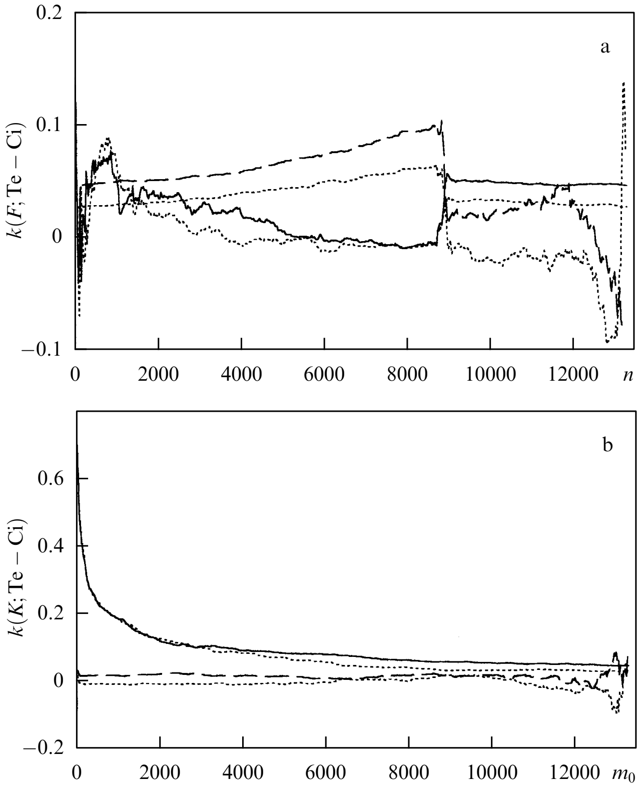

Figure 20. Current correlation coefficients for the method of sweeps in the structural spectra for Te–Ci correlations in the collagen gene. The solid curves describe the inflation from left to right, while the dashed curves describe the inflation from right to left. Shown by the dotted lines are the corresponding dependences obtained after random reshuffling of exons and introns while preserving their alternation.

dependence for $k(K;\ \mathrm{Te-Ci})$ obtained by sweeping from left to right, we observe a rapid decrease in correlations for $m_0 \geqslant 10^3$ and relatively weak residual large-scale correlations, which are clearly shown by a dashed line corresponding to the inflation of the spectrum for correlation functions from right to left (Fig. 20b). When the exons and the introns are randomly reshuffled while preserving their alternation, we observe that the residual large-scale correlations at $m_0 \geqslant 10^3$ are destroyed and the integral correlation coefficient is decreased by a factor of 1.5 – 2 (dotted lines in Fig. 20). A random reshuffling of introns or exons alone gives a similar effect. The dependences shown in Fig. 20 indicate the existence of the hidden periodicity $p = 3$ in some introns (direct analysis does reveal this periodicity in several introns). Such intron fragments could dynamically bond to mRNA thereby affecting the rate of protein synthesis [126].

The analysis of correlations in the different exon – intron sequences shows that there exists a variety of mechanisms for structural coupling [27]. In addition to the coupling via the periodicity $p = 3$, in the correlations there may be a contributions of $p = 2$ and other periodicities as well; the shifts between the periodicities in exons and introns can also result in anticorrelations (in this case the jumps similar those in Fig. 20 are negative). In some cases, on the background of a slow decrease of correlations on scales $m_0 \sim 10^3$, on smaller scales $m_0 \sim 10^2$ we observe the characteristic correlations indicating the stronger similarity between the structural characteristics for some shorter segments.

The exact removal of introns during splicing is ensured by a specific recognition of the boundaries between exons and introns [29, 31]. At the boundary between the 3$'$ end of an exon and the 5$'$ end of an intron there is a consensus sequence ${}^{\mathrm{A}}_{\mathrm{C}}$AG|GT${}^{\mathrm{A}}_{\mathrm{G}}$AGT, while the corresponding sequence for the boundary between the 3$'$ end of an intron and the 5$'$ end of an exon is $(\mathrm{Y})_{11}$NYAG|N, where Y denotes a pyrimidine, N is any nucleotide, and the vertical line indicates the position of the boundary. The characteristic lengths of the approximately conserved consensus fragments are about 10 – 30 nucleotides. However, we observe that the range of structural coupling between exons and introns is much greater. The couplings may be related to the packing of DNA into nucleosomes, splicing mechanisms, regulation of protein synthesis, etc. [37, 120, 126, 26, 27].

The examples discussed in Sections 5.1 and 5.2 show that the joint use of spectra for the structure factors and correlation functions allows one to reveal a set of characteristic lengths and to determine the mechanisms responsible for the correlations observed. In addition, the spectral methods appear to be practically unique among all known methods allowing us to reveal and estimate quantitatively the significance of rather weak effects of structural coupling between different elements of the mosaic structure of a genome.

6. Conclusions

The one-to-one correspondence [see (3.2) and (3.3)] and mutual complementarity of scales, $q_n \propto m^{-1}$, m_0^{-1} [see (3.3) and (3.6)] for the Fourier transform provide a complete representation of the structural characteristics for DNA sequences. Unlike, for example, the Markov chain theory (Section 2.2), the spectral approach is not based on *a priori* assumptions concerned with the character of regularity. Moreover, the character of regularity and the nature of mutual correlations are revealed during spectral analysis based on the joint use of structure factors (3.5) and correlation functions (3.6). The integral regularity of a DNA sequence can be estimated in terms of the structural entropy (3.44), (3.47). Spectral analysis reveals a wide range of characteristic lengths, some of which are related to segmentation effects. If necessary, segmentation effects can be taken into account from the very beginning (Section 5.2).

The local variations in the structural characteristics can be studied with the use of the standard technique with windows sliding along a sequence [5, 7, 8]. Using the methods of filtration theory [91], we can filter out different characteristic regions in the Fourier spectra, and solve an inverse problem to find the structural features responsible for the singular harmonics (cf. [19]). Such an approach can be useful in preliminary studies of the segmentation, while the exact positions of segment boundaries can be determined with the use of a catalogue of the specific signal sequences.

In contrast with a purely statistical approach based on the calculation of occurrence frequencies for different nucleotide combinations [5 – 9], spectral methods allow one to reveal a number of reproducible structural features in DNA segments with different functions. Indeed, in Ref. [45] the efficiency was studied of more than twenty methods for the identification of protein-coding regions. For windows of width $w = 120$ the Fourier method appears to be the most efficient. It is followed by the method based on the calculation of hexanucleotide combinations. In the latter method we have to analyze as many as $4^6 = 4096$ values, while using the Fourier method we can consider only a few harmonics in the vicinity of the periodicity $p = 3$. It is essential that upon revealing reproducible structural features, we can understand the general principles underlying the formation of genomic DNA sequences and, as a result, obtain a basis for the development of various evolutionary theories [4, 41, 96, 97]. From that viewpoint, the DNA is a unique object; on the one hand, DNA contains information on 'modern' proteins, which are interesting for medical and biochemical applications [29] or even for development of new nanotechnologies [127], on the other hand, the DNA sequences can be considered as the annals of molecular evolution over about 4 billion years [4, 6, 9, 29, 41, 96, 97].

When analyzing all the sequences in a database, we should remember that the initial stage of data accumulation is not over yet. Only in the near future shall we obtain the complete sequences for genomes of higher organisms. In many respects the set of sequences available nowadays in databases is dictated by medical, biochemical, and some other applications. The set is rather specific and is still far from being statistically homogeneous. Therefore, the comparative analysis of different species and evolutionary studies require a careful preliminary selection of the analogous data [4, 6, 8, 9, 29, 128]. We should also remember that DNA is organized in a complicated hierarchical system, where each level is related to the corresponding genetic and biological mechanisms (and in this sense DNA is essentially different from homogeneous fractals [125]). Based only on statistical considerations the estimation of DNA sequence complexity without preliminary reconstruction of the hierarchy may appear to be similar to estimating electronic circuit complexity by the length of connecting wires.

Within the framework of the spectral approach, it is possible to attain a formal unification of the digital representation of the data and similarity in statistics for random analogues of the spatial characteristics for protein C_α-backbones, distributions of physico-chemical parameters along a polypeptide chain, and the structural characteristics of the corresponding protein-coding DNA sequences [24, 129]. This gives practically a unique opportunity for studying the direct correlations between these values. The general mathematical theory for structural analysis of spatial linear chains within the framework of the spectral approach is outlined in [130]. Using the X-ray crystallography data, the relationship between the spatial structures of C_α-backbones of proteins and the distribution of physico-chemical parameters along the protein chains was studied in Ref. [129], and it was qualitatively shown in [24] that the $\Delta S_{\rm rel}$ for protein-coding DNA depends on the regularity of the spatial protein structure. A detailed study of the relationships between all these characteristics is of fundamental interest and would also allow one to understand how the spatial structure of a protein molecule manifests itself in the protein-coding DNA sequences.

The results of the structural analysis confirm a certain structural integrity of the genomic DNA sequences. Despite the huge material collected, in many respects the general principles of genomic sequence formation are not clearly understood yet. The understanding of the language of genetic texts is far from complete. The ordering and the selection of reproducible features should be followed by the next stage of investigation concerned with the physical mechanisms responsible for the structural features observed.

Finally, the authors acknowledge A A Vedenov for support of the research program. We are thankful to A M Kamchatnov, E B Levchenko, M A Lifshits, V Yu Makeev, and A L Chernyakov for fruitful discussions of various aspects of the problems concerned, and also to A Yu Turygin for assistance in performing numerical calculations and to A A Ezhov who took part in developing the methods for reconstruction of hidden repeats. We are also thankful to our numerous colleagues for delivering their results, frequently prior to publication.

References

1. Rodriguez-Tome P et al. *Nucl. Acids Res.* **24** 6 (1996)
2. Benson D A et al. *Nucl. Acids Res.* **26** 1 (1998)
3. Stoesser G et al. *Nucl. Acids Res.* **27** 18 (1999)
4. Ratner V A et al. *Problemy Teorii Molekulyarnoĭ Evolutsii* (Problems in Theory of Molecular Evolution) (Novosibirsk: Nauka, 1985) [Translated into English: *Molecular Evolution* (Berlin: Springer, 1996)]
5. Aleksandrov A A et al. *Komp'yuternyĭ Analiz Geneticheskikh Tekstov* (Computer Analysis of Genetic Texts) (Ed. M D Frank-Kamenetskiĭ) (Moscow: Nauka, 1990)
6. Weir B S *Genetic Data Analysis* (Sunderland, Mass.: Sinauer Associates, Inc. Publ., 1990) [Translated into Russian (Moscow: Mir, 1995)]
7. Waterman M S (Ed.) *Mathematical Methods for DNA Sequences* (Boca Raton, Fl.: CRC Press, 1989) [Translated into Russian (Moscow: Mir, 1999)]
8. Bishop M J, Rawlings C J (Eds) *DNA and Protein Sequence Analysis*: *A Practical Approach* (Oxford: IRL Press, 1997)
9. Kolchanov N A, Lim H A *Computer Analysis of Genetic Macromolecules*: *Structure, Function and Evolution* (Singapore: World Scientific, 1994)
10. Swindell S R, Miller R R, Myers G S A (Eds) *Internet for the Molecular Biologist* (Oxford: Horizon Scientific, 1996)
11. Baxevanis A, Quellette B F F *Bioinformatics*: *A Practical Guide to the Analyses of Genes and Proteins* (New York: Wiley, 1998)
12. Vingron M, Waterman M S *J. Mol. Biol.* **235** 1 (1994)
13. Gelfand M S *J. Comp. Biol.* **2** 87 (1995); Gel'fand M S *Mol. Biol.* **32** 103 (1998)
14. Fickett J W *Comput. Chem.* **20** 103 (1996)
15. Li W *Comput. Chem.* **21** 257 (1997)
16. http://linkage.rockefeller.edu/wli/dna_corr
17. Turygin A Yu, Chechetkin V R *Zh. Eksp. Teor. Fiz.* **106** 335 (1994) [*JETP* **79** 176 (1994)]
18. Chechetkin V R, Turygin A Y *J. Phys. A* **27** 4875 (1994)
19. Chechetkin V R, Knizhnikova L A, Turygin A Y *J. Biomol. Struct. Dyn.* **12** 271 (1994)
20. Chechetkin V R, Turygin A Y *Phys. Lett. A* **199** 75 (1995)
21. Chechetkin V R, Turygin A Y *J. Theor. Biol.* **175** 477 (1995)
22. Chechetkin V R et al. "Reconstruction of hidden repeats in DNA sequences, their significance, and applications", Preprint TRINITI 0033-A (Moscow: TsNIIATOMINFORM, 1996) [in English]
23. Chechetkin V R, Turygin A Y *J. Theor. Biol.* **178** 205 (1996)
24. Chechetkin V R, Lobzin V V *Phys. Lett. A* **222** 354 (1996)
25. Ezhov A A, Chechetkin V R *Mat. Modelirovanie* **10** 83 (1998)
26. Chechetkin V R, Lobzin V V *J. Biomol. Struct. Dyn.* **15** 937 (1998)
27. Chechetkin V R, Lobzin V V *J. Theor. Biol.* **190** 69 (1998)
28. Ivanitskiĭ G R et al. *Biofiz.* **30** 418 (1985)
29. Alberts B et al. *Molecular Biology of the Cell* (New York: Garland Publ., 1983) [Translated into Russian (Moscow: Mir, 1986)]
30. Rees A R, Sternberg M J E *From Cells to Atoms. An Illustrated Introduction to Molecular Biology* (Oxford: Blackwell, 1984) [Translated into Russian (Moscow: Mir, 1988)]
31. Georgiev G P *Geny Vysshikh Organizmov i Ikh Ekspressiya* (Genes of Higher Organisms and Its Expression) (Moscow: Nauka, 1989)
32. Grosberg A Yu, Khokhlov A R *Statisticheskaya Fizika Makromolekul* (Statistical Physics of Macromolecules) (Moscow: Nauka, 1989) [Translated into English (New York: AIP Press, 1994)]
33. Creighton T *Proteins*: *Structures and Molecular Properties* (New York: Freeman, 1993)
34. Stepanov V M *Struktura i Funktsii Belkov* (Structure and Functions of Proteins) (Moscow: Vysshaya Shkola, 1996)
35. Britten R J, Kohne E D *Science* **161** 529 (1968)
36. Trifonov E N *Bull. Math. Biol.* **51** 417 (1989)
37. Trifonov E N *Comput. Chem.* **17** 27 (1993)
38. Trifonov E N, Sussman J L *Proc. Natl. Acad. Sci. USA* **77** 3816 (1980)
39. Hagerman P J *Annu. Rev. Biochem.* **59** 755 (1990)
40. Olson W K, Zhurkin V B, in *Biological Structure and Dynamics* (Eds R H Sarma, M H Sarma) (New York: Adenine Press, 1996) p. 341
41. Nussinov R *J. Theor. Biol.* **125** 219 (1987)
42. Staden R *Comput. Applic. Biosci.* **4** 53 (1988)
43. Sprizhitskiĭ Yu A et al. *Mol. Biol.* **22** 338 (1988)
44. Wada K et al. *Nucl. Acids Res.* **19** 1981 (1991)
45. Fickett J W, Tung C-S *Nucl. Acids Res.* **20** 6441 (1992)
46. Shannon C E *Raboty po Teorii Informatsii i Kibernetike* (Papers on Information Theory and Cybernetics) (Moscow: IIL, 1963) [see: Claude Elwood Shannon: Collected Papers (Eds W J A Sloane, A D Wyner) (New York: IEEE Press, 1993)]
47. Bleisdell B E *J. Mol. Evol.* **21** 278 (1985)
48. Borodovskiĭ M Yu et al. *Mol. Biol.* **20** 1014, 1024, 1390 (1986)
49. Phillips G J, Arnold J, Ivarie R *Nucl. Acids Res.* **15** 2611, 2627 (1987)
50. Stückle E E et al. *Nucl. Acids Res.* **18** 6641 (1990)
51. Feller W *An Introduction to Probability Theory and Its Applications* (New York: Wiley, 1970) [Translated into Russian (Moscow: Mir, 1984)
52. Herzel H, Ebeling W, Schmitt A O *Phys. Rev. E* **50** 5061 (1994); *Chaos, Solitons and Fractals* **4** 97 (1994)
53. Mantegna R N et al. *Phys. Rev. Lett.* **73** 316 (1994)
54. Israeloff N E, Kagalenko M, Chan K *Phys. Rev. Lett.* **76** 1976 (1996)
55. Schmitt A O, Ebeling W, Herzel H *Biosystems* **37** 199 (1996)
56. Chatzidimitriou-Dreismann C A, Streffer R M F, Larhammar D *Nucl. Acids Res.* **24** 1676 (1996)
57. Luo L F *Collected Works on Theoretical Biophysics* (Hohhot: Inner Mongolia University Press, 1997)
58. Herzel H, Grosse I *Physica A* **216** 519 (1995)
59. Mrazek J, Kypr J *Comp. Applic. Biosci.* **11** 195 (1995)
60. Li W, Kaneko K *Europhys. Lett.* **17** 655 (1992)
61. Luo L F et al. *Phys. Rev. E* **58** 861 (1998)
62. Peng C-K et al. *Nature* (London) **356** 168 (1992)

63. Nee S *Nature* (London) **357** 450 (1992)
64. Karlin S, Brendel V *Science* **259** 677 (1993)
65. Chatzidimitriou-Dreismann C A, Streffer R M F, Larhammar D *Eur. J. Biochem.* **224** 365 (1994); *Biochim. Biophys. Acta* **1217** 181 (1994)
66. Kapitonov V V, Titov I I *Dokl. Ross. Akad. Nauk* **337** 810 (1994)
67. Borovik A S, Grosberg A Y, Frank-Kamenetskii D F *J. Biomol. Struct. Dyn.* **12** 655 (1994)
68. Shnerb N, Eisenberg E *Phys. Rev. E* **49** R1005 (1994)
69. Stanley H E et al. *Nuovo Cimento D* **16** 1339 (1996)
70. Allegrini P et al. *Phys. Rev. E* **57** 4558; **58** 3640 (1998)
71. Viswanathan G M et al. *Physica A* **249** 581 (1998)
72. Astaf'eva N M *Usp. Fiz. Nauk* **166** 1145 (1996) [*Phys. Usp.* **39** 1085 (1996)]
73. Arneodo A et al. *Phys. Rev. Lett.* **74** 3293 (1995)
74. Altaiskii M, Mornev O, Polozov R *Genetic Analysis*: *Biomol. Engineer.* **12** 165 (1996)
75. Tsonis A A et al. *Phys. Rev. E* **53** 1828 (1996)
76. Arneodo A et al. *Physica A* **249** 439 (1998)
77. Bernardi G *Annu. Rev. Genet.* **29** 445 (1995)
78. Fickett J W, Torney D C, Wolf D R *Genomics* **13** 1056 (1992)
79. McLachlan A D, Stewart M *J. Mol. Biol.* **103** 271 (1976)
80. Silverman B D, Linsker R *J. Theor. Biol.* **118** 295 (1986)
81. Deev A A et al. *Biofiz.* **34** 564 (1989)
82. McLachlan A D *J. Phys. Chem.* **97** 300 (1993)
83. Makeev V J, Tumanyan V G *Comp. Applic. Biosci.* **12** 49 (1995)
84. Lee W J, Luo L F *Phys. Rev. E* **56** 848 (1997)
85. Kutuzova G I et al. *Biofiz.* **42** 354 (1997)
86. Felsenstein J, Sawyer S, Kochin R *Nucl. Acids Res.* **10** 133 (1982)
87. Benson D C *Nucl. Acids Res.* **18** 3001, 6305 (1990)
88. Cheever E A, Overton G C, Searls D B *Comp. Applic. Biosci.* **7** 143 (1991)
89. Arques D G, Michel C J, Oriex K *Comp. Applic. Biosci.* **8** 5 (1992)
90. Voss R F *Phys. Rev. Lett.* **68** 3805 (1992); *Fractals* **2** 1 (1994)
91. Marpl S L (Jr) *Digital Spectral Analysis with Applications* (Englewood Cliffs, N.J.: Prentice-Hall, 1987) [Translated into Russian (Moscow: Mir, 1990)]
92. Van Kampen N G *Stochastic Processes in Physics and Chemistry* (Amsterdam: North-Holland, 1984) [Translated into Russian (Moscow: Vysshaya Shkola, 1990)]
93. Anderson T W *An Introduction to Multivariate Statistical Analysis* (New York: Wiley, 1958) [Translated into Russian (Moscow: Fizmatgiz, 1963)
94. Low R L, Arai K, Kornberg A *Proc. Natl. Acad. Sci. USA* **78** 1436 (1981)
95. Green N P O, Stout G W, Taylor D J *Biology* (Cambridge: Cambridge University Press, 1985) [Translated into Russian (Moscow: Mir, 1993)]
96. Crick F H C et al. *Origins Life* **7** 389 (1976)
97. Eigen M et al. *Sci. Am.* **244** 88 (1981)
98. Zhurkin V B *Nucl. Acids Res.* **9** 1963 (1981)
99. Kypr J, Mrazek J *Int. J. Biol. Macromol.* **9** 49 (1987)
100. Leadbetter M R, Lindgren G, Rootzen H *Extremes and Related Properties of Random Sequences and Processes* (New York: Springer, 1983)
101. Mani G A *J. Theor. Biol.* **158** 447 (1992)
102. Eigen M, Winkler-Oswatitsch R *Naturwissenschaften* **68** 282 (1981)
103. Lagunez-Otero J, Trifonov E N *J. Biomol. Struct. Dyn.* **10** 451 (1992)
104. Curran J F, Gross B L *J. Mol. Biol.* **235** 389 (1994)
105. Almagor H *J. Theor. Biol.* **117** 127 (1985)
106. Zhang C-T, Chou K-C *J. Mol. Biol.* **238** 1 (1994)
107. Mrazek J, Kypr J *J. Mol. Evol.* **39** 439 (1994)
108. Lio P, Ruffo S, Buiatti M *J. Theor. Biol.* **171** 215 (1994)
109. Frank G K, Makeev V J *J. Biomol. Struct. Dyn.* **14** 629 (1997)
110. Johnson R L et al. *Science* **272** 1668 (1996)
111. Gailani M R et al. *Nature Genetics* **14** 78 (1996)
112. Ezhov A A et al. *Ross. Zh. Kozhnykh Bolezneĭ* (1) 17 (1999)
113. Godreche C, Luck J M *J. Phys. A* **23** 3769 (1990)
114. Cheng Z, Savit R *Phys. Rev. A* **44** 6379 (1991)
115. Li W *Phys. Rev. A* **43** 5240 (1991)
116. Wang Y-H et al. *Science* **265** 669 (1994)
117. Bates G, Lehrach H *Bioessays* **16** 277 (1994)
118. *Analysis of Triplet Repeat Disorders* (Eds D Rubinsztein, M Hayden) (Oxford: BIOS, 1998)
119. Bina M *J. Mol. Biol.* **235** 198 (1994)
120. Beckmann J S, Trifonov E N *Proc. Natl. Acad. Sci. USA* **88** 2380 (1991)
121. Arques D G, Michel C J *Nucl. Acids Res.* **15** 7581 (1987)
122. Vogt P *Human Genet.* **84** 301 (1990)
123. Arques D G, Michel C J *J. Theor. Biol.* **143** 307 (1990)
124. Hamermesh M *Group Theory and Its Applications to Physical Problems* (Reading, Mass.: Addison-Wesley, 1962) [Translated into Russian (Moscow: Mir, 1966)]
125. Feder J *Fractals* (New York: Plenum, 1988) [Translated into Russian (Moscow: Mir, 1991)]
126. Novak R *Science* **263** 608 (1994)
127. Drexler K E *Annu. Rev. Biophys. Biomol. Struct.* **23** 377 (1994)
128. Ayala F J *Population and Evolutionary Genetics*: *A Primer* (Reading, Mass.: Benjamin Cummings Reading, 1982) [Translated into Russian (Moscow: Mir, 1984)]
129. Chechetkin V R, Lobzin V V *J. Theor. Biol.* **198** 197, 219 (1999)
130. Lobzin V V, Chechetkin V R *Zh. Eksp. Teor. Fyz.* **116** 620 (1999) [*JETP* **89** 331 (1999)]