DNA Research **18**, 353–362, (2011)
Advance Access Publication on 25 July 2011
doi:10.1093/dnares/dsr023

Profile-Statistical Periodicity of DNA Coding Regions

Maria Chaley[1,*], and Vladimir Kutyrkin[2]

Institute of Mathematical Problems of Biology, Russian Academy of Sciences, Institutskaya st., 4, 142290 Pushchino, Russia[1] and Department of Computational Mathematics and Mathematical Physics, Moscow State Technical University n.a. N.E. Bauman, the 2nd Baumanskaya st., 5, 105005 Moscow, Russia[2]

*To whom correspondence should be addressed. Tel. +7 4967-318537. Fax. +7 4967-318500.
E-mail: maramaria@yandex.ru

Edited by Hiroyuki Toh
(Received 7 March 2011; accepted 13 June 2011)

Abstract

Novel methods for identifying a new type of DNA latent periodicity, called latent profile periodicity or latent profility, are used to search for periodic structures in genes. These methods reveal two distinct levels of organization of genetic information encoding. It is shown that latent profility in genes may correlate with specific structural features of their encoded proteins.

Key words: latent periodicity; latent profility; spectral–statistical approach; gene structure

1. Introduction

The notion of latent periodicity in nucleotide DNA sequences has arisen from the discovery of various regularity levels in the structural organization of a DNA molecule. For example, a DNA double-helix pitch has been found to equal ~10–11 bp, a length of ~200 bp has been found for DNA fragments in nucleosomes, and a loop length of $\sim 2 \times 10^4$–10^5 bp has been determined at the higher level of quasi-regular DNA compactization.[1] Such particularities are probably due to non-random alternation of bases in the original DNA sequence. Thus, research on both short- and long-range base correlations[2] is of great importance for understanding known structural particularities in DNA sequences and for revealing new ones.

Images of various functions demonstrating nucleotide correlations in DNA coding regions show regular peaks with the steps of three bases corresponding to the triplet nature of the genetic code. This has led to the notion of triplet periodicity in the coding regions. It is likely that use of the term 'latent triplet periodicity' has also been influenced by the hypothesis on the universal RNY triplet pattern (R, purine; Y, pyrimidine; N, purine or pyrimidine) of codons in ancient genes, because correlations with this pattern can be traced in base distribution over recent coding regions.[3]

In light of the current understanding of latent periodicity as approximate tandem repeats,[4] the occurrence of periodicity has been corroborated by textual 'consensus pattern', which is an estimate of a pattern in the original repeat. If alterations in the copies of the pattern account for more than 30% of the pattern, the validity of the revealed consensus pattern is in doubt. Although tandem repeats of tri- and hexa-nucleotides occur in coding regions,[5,6] as a rule it is impossible to deduce a reliable consensus pattern of an approximate tandem repeat over the whole length of a coding region.

Weak three-base periodicity in *Escherichia coli* mRNA with pattern G-non-G-N, complemented by pattern NNC in 16S RNA, has been considered as a mechanism for monitoring translation frames in ribosomes.[7] Despite the discovery of a few short tracks of corresponding patterns, this case would be more accurate to consider only as the domination of G and C bases in corresponding triplet positions. The weak preference of a certain type of base in a fixed

position of triplets in the coding region promotes the appearance of a dominant peak in the Fourier spectrum at a frequency of 0.33 corresponding to the three-base period.[8] Nevertheless, this preference is not the key cause behind this observation. It appears that the greater the variance of certain base distributions over period positions, the more impact such a base has on the amplitude of spectral density at the three-period frequency, even when the base is not dominant at triplet positions.[9] Thus, appearance of the peak at the frequency of 0.33 in the Fourier spectrum is due to non-uniform base distribution over the triplet positions.

Use of the Fourier methods for revealing imperfect periodicity[8,10–13] has become common. Other statistical methods have also arisen for determining latent periodicity in nucleotide sequences.[14–16] These methods are based on measuring heterogeneity in base distributions over period positions.[14] In practice, in the absence of weak periodicity in a sequence that is not an approximate tandem repeat, a high index of heterogeneity and a Fourier spectrum with a dominant peak may be observed. It is incorrect in this case to use the term 'latent periodicity', until the discovery of a pattern indicating some new type of periodicity, for example, such as the flexible patterns[17] of 11th nucleotide periodicity in the genomes of prokaryotes and low eukaryotes.

In the present work, a spectral–statistical approach (2S approach)[14,18] to identifying a new type of latent periodicity, called profile periodicity or profility, is proposed. The notion of latent profility in DNA sequences has been introduced earlier.[14] It expands on the idea of approximate tandem repeat[4] in which textual string (DNA sequence) is presented as a chain of eroded copies (with $\sim$80% identity) of the textual pattern. Latent profile periodicity occurs in DNA regions where nucleotide correlations can be described by hypothesizing on the generation of successive uni-length DNA fragments according to a fixed probability distribution of bases appearance at each fragment position. A pattern of latent profility can be described with the aid of a finite random string consisting of independent random characters with corresponding probability distribution for the textual characters from the DNA alphabet. Excepting certain cases, the DNA sequences considered in this work could not be identified as approximate tandem repeats. The usual methods[4,19–22] applied for identification of approximate tandem repeats therefore cannot be used to reveal new types of latent periodicity.

Via a procedure that identifies patterns of latent profile periodicity in DNA, it is possible to reveal the two levels of organization in encoded genetic information: regular heterogeneity of nucleotide distribution over positions in the codons and latent profility. It is shown that the Fourier analysis does not enable the second level (latent profility) to be distinguished in the genetic encoding. The Fourier spectra for the DNA coding regions have been built using FFT programs,[12,13] available from http://www.imtech.res.in/raghava/ftg/.

A number of examples are given when the latent profility revealed in the DNA coding regions is translated into features of protein structure. The direct detection of such features is challenging because the goal of the search is not *a priori* known. The present work shows that it is possible to determine signs of local features in the structure of genes and corresponding proteins through latent profile periodicity (latent profility).

2. Methods of identifying latent profile periodicity in DNA

In the proposed model of latent profile periodicity, a DNA sequence (textual string) is considered as a realization of a special random periodic string called a profile string. This string, consisting of independent random characters, represents a perfect tandem repeat of a random string, so called because of its random periodicity pattern. Therefore, in order to determine the latent profile periodicity (profility) in the DNA sequence (textual string), it is necessary to specify the criteria for considering an analysed textual string as a realization of some profile string. The random pattern of such a profile string is not dependent on the quality of the periodicity pattern for the analysed textual string.

2.1. *Statistical structure of random strings consisting of independent random characters*

DNA sequences are considered as textual strings in the four-character ($K = 4$) simply ordered alphabet $A = \langle a, g, t, c\rangle$, where a is the adenine, g the guanine, t the thymine and c the cytosine.

Let $\mathrm{Chr}(\mathbf{p})$ be a random character with the frequency column $\mathbf{p} = (p^1, \ldots, p^K)^{\mathrm{T}}$. Such a character is a random variable that takes the value of the ith character of the alphabet $A = \langle a_1, \ldots, a_K\rangle$ with a probability of p^i $(i = 1, \ldots, K)$.

A special random string $\mathrm{Str}_n(\boldsymbol{\pi}) = \mathrm{Chr}(\mathbf{p}_1)\cdots\mathrm{Chr}(\mathbf{p}_n)$ of n independent random characters is induced by the matrix $\boldsymbol{\pi} = (\mathbf{p}_1, \ldots, \mathbf{p}_n) = (\pi_j^i)_n^K$, called an n-profile matrix. Let $\mathrm{str} = a_{i_1}\cdots a_{i_n}$ be a textual string, where i_m is the number of a character a_{i_m} $(m = 1, \ldots, n)$ in alphabet A. If the str is a realization of the random string $\mathrm{Str}_n(\boldsymbol{\pi})$, then the product $\pi_1^{i_1}\cdots\pi_n^{i_n}$ determines the probability of such a realization.

The character $a_i \in A (i = 1, \ldots, K)$ can be identified with a random character for which all components of the frequency column are null, excepting for the ith, which is a unity component. Therefore, any textual string in alphabet A can be identified with the corresponding special random string of the same length.

An integer number L from the diapason $1, \ldots, L_{\max}$, where $L_{\max} \sim n/5K$, is called the test-period of the string $\mathrm{Str}_n(\boldsymbol{\pi})$.

Let L be the test-period of random string $\mathrm{Str} = \mathrm{Str}_n(\boldsymbol{\pi})$, and $\mathrm{Str}_n(\boldsymbol{\pi}) = \mathrm{Str}_L(\boldsymbol{\pi}_1) \cdots \mathrm{Str}_L(\boldsymbol{\pi}_m)\, \mathrm{Str}_M(\boldsymbol{\pi}_{m+1})$ be a decomposition of string Str into substrings of length L, where $0 \le M < L$ (if $M \neq 0$, substring $\mathrm{Str}_M(\boldsymbol{\pi}_{m+1})$ is not complete; if $M = 0$, substring $\mathrm{Str}_M(\boldsymbol{\pi}_{m+1})$ is empty). Then, if $M = 0$, the matrix $\Pi_{\mathrm{Str}}(L) = (1/m) \sum_{i=1}^{m} \boldsymbol{\pi}_i$ is called the L-profile matrix of string Str. If $M \neq 0$, then corrections are made in the $\Pi_{\mathrm{Str}}(L)$ matrix. Thus, the profile-matrix spectrum Π_{Str}, defined at each test-period, is introduced for string Str. The profile-matrix spectrum characterizes the statistical structure of the realizations of random string Str. If the statistical structures of string Str and of analysed textual string str are indistinguishable (at a corresponding level of significance), then it can be considered that the string str is a realization of random string Str. Further methods for verifying this will be proposed on the basis of the latent profile periodicity model.

2.2. A stochastic model of the latent profile periodicity

Occurrence of the latent profile periodicity in the analysed textual string manifests in the statistical string structure observed in the sample profile-matrix spectrum. In fact, if the analysed string is sufficiently long, this sample spectrum takes the form of the profile-matrix spectrum of periodic random string Str consisting of independent random characters. In this case, random string Str is given by $\mathrm{Str} = \mathrm{Str}_L(\boldsymbol{\pi}_1) \cdots \mathrm{Str}_L(\boldsymbol{\pi}_m)\mathrm{Str}_M(\boldsymbol{\pi}_{m+1})$, where L is the period of string Str, $0 \le M < L$, $\boldsymbol{\pi}_1 = \cdots = \boldsymbol{\pi}_m = \boldsymbol{\pi}_0$ and $\mathrm{Str}_L(\boldsymbol{\pi}_0) = \mathrm{Str}_M(\boldsymbol{\pi}_{m+1})\mathrm{Str}_{L-M}(\boldsymbol{\pi}_{10})$. Such a string Str is called the L-profile string with a random periodicity pattern $\mathrm{Str}_L(\boldsymbol{\pi}_0)$. Moreover, in this case, the designation $\mathrm{Tdm}_L(\boldsymbol{\pi}_0, n)$ is used for the string Str.

Matrix $\boldsymbol{\pi}_0$ is called the main profile matrix of the string $\mathrm{Str} = \mathrm{Tdm}_L(\boldsymbol{\pi}_0, n)$ because matrix $\boldsymbol{\pi}_0$ initiates the entire profile-matrix spectrum of this string. Profile string $\mathrm{Tdm}_L(\boldsymbol{\pi}_0, n)$ is a perfect tandem repeat with a random periodicity pattern $\mathrm{Str}_L(\boldsymbol{\pi}_0)$. The profile-matrix spectrum of the string $\mathrm{Tdm}_L(\boldsymbol{\pi}_0, n)$ can be considered as a stochastic model of heterogeneity manifestation in textual strings that are realizations of the string $\mathrm{Tdm}_L(\boldsymbol{\pi}_0, n)$.

2.3. Size estimation for pattern of latent profile periodicity in a textual string

To estimate the pattern size of the latent profile periodicity, a characteristic spectrum is established for a textual string. Characteristic spectra of the three approximate tandem repeats from the database TRDB (http://tandem.bu.edu/cgi-bin/trdb/) are shown in Fig. 1a–c. In each of these spectra, the first clear maximum arises at the test-period that is a period of approximate tandem repeat. A similar observation is made for the characteristic spectra of textual strings with latent profile periodicity (Fig. 2a–c). Therefore, the first test-period at which the maximum value of the characteristic spectrum for the analysed textual string is clear is used to estimate the pattern size for the latent profile periodicity, or profility.

The characteristic spectrum for the analysed textual string str of length n in alphabet A is determined as follows. For every test-period Λ of this string, the profile string $\mathrm{Tdm}_\Lambda = \mathrm{Tdm}_\Lambda(\Pi_{\mathrm{str}}(\Lambda), n)$ is created, and the Pearson statistics[14] is introduced:

$$\psi(\Pi_{\mathrm{str}}(\lambda), \Pi_{\mathrm{Tdm}_\Lambda}(\lambda), n) = \frac{n}{\lambda} \sum_{j=1}^{\lambda} \sum_{i=1}^{K} \frac{(\pi *_j^{\,i} - \pi_j^i)^2}{\pi_j^i (1 - \pi_j^i)} \sim \chi^2_{(K-1)\lambda}, \tag{1}$$

where $\Pi_{\mathrm{str}}(\lambda) = (\pi *_j^{\,i})_\lambda^K$ and $\Pi_{\mathrm{Tdm}_\Lambda}(\lambda) = (\pi_j^i)_\lambda^K$, and χ^2_N is a $\chi 2$ distribution with N degrees of freedom. When $\Lambda = 1$, the value of the characteristic spectrum $H(\lambda)$ at the test-period λ is calculated by

$$H(\lambda) = \psi(\Pi_{\mathrm{str}}(\lambda),\, \Pi_{\mathrm{Tdm}_1}(\lambda), n) - E(\chi^2_{(K-1)\lambda}), \tag{2}$$

where $E(\chi^2_N)$ is the mathematical expectation of χ^2_N.

As noted above, the first test-period L with a clear maximum value of spectrum H provides an estimate of the latent period of profility in string str (Fig. 2a–c).

2.4. Estimation of pattern of latent profile periodicity

Let L be the proposed estimation of pattern size for the latent profile periodicity of analysed textual string str of length n in alphabet A of K characters. Then, to estimate the pattern of the latent periodicity in this string, the periodicity pattern of profile string $\mathrm{Tdm}_L = \mathrm{Tdm}_L(\Pi_{\mathrm{str}}(L), n)$ is proposed. Hence, $\mathrm{Str}_L(\Pi_{\mathrm{str}}(L))$ is an estimation of the pattern of latent profile periodicity in an analysed string str. If such an estimation is valid then string str is statistically indistinguishable from the profile string Tdm_L. In this case, string str can be considered a realization of the string Tdm_L.

To check the statistical indistinguishability of strings str and Tdm_L, the D_L spectrum, called a spectrum of

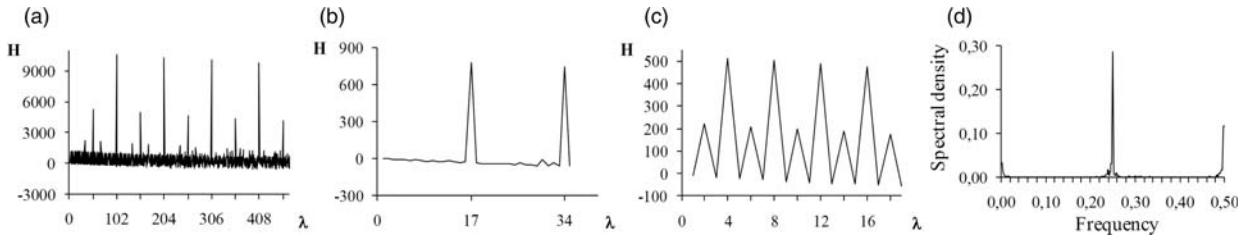

Figure 1. Characteristic spectra of approximate tandem repeats from the TRDB database for (a) *C. elegance* chromosome IV (688 744–698 236 bp, period = 102 bp, %mismatches = 3, %indels = 0, copies = 93.1), (b) *C. elegance* chromosome V (1 809 784–1 810 492 bp, period = 17 bp, %mismatches = 12, %indels = 0, copies = 41.8), (c) *M. musculus* chromosome I (26 399 024– 26 399 410 bp, period = 12 bp, %mismatches = 17, %indels = 0, copies = 32.5). (d) Fourier spectrum for the tandem repeat (c) of *M. musculus*. The maximal peak is reached at frequency 0.25 and corresponds to the period of 4 bp.

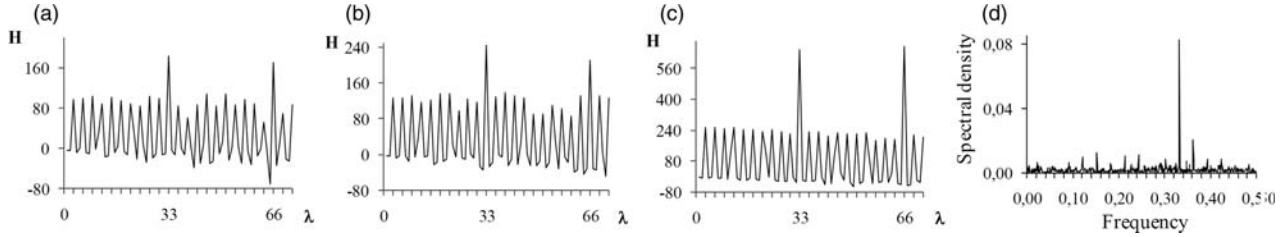

Figure 2. Characteristic spectra of mRNA coding regions of the PF01442 apolipoprotein family from the Pfam database for (a) *M. musculus* Apo E (GenBank M12414, region: 1–936 bp), (b) *S. aurata* Apo A-I (GenBank AF013120, region: 34–816 bp), (c) *G. gallus* Apo A-IV (GenBank Y16534, region: 37–1137 bp). (d) Fourier spectrum for the coding region of the apolipoprotein mRNA (c) of *G. gallus*. The maximal peak is reached at frequency 0.33 and corresponds to 3-regularity.

string str deviation from L-profility, is used. At test-period λ of string str the D_L spectrum has the value

$$D_L(\lambda) = \frac{\psi(\Pi_{\text{str}}(\lambda), \Pi_{\text{Tdm}_L}(\lambda), n)}{\chi^2_{\text{crit}}((K-1)\lambda, \alpha)} \qquad (3)$$

where ψ statistics has been introduced in Equation (1), and $\chi^2_{\text{crit}}(N, \alpha)$ is the left-hand-side critical value of the χ^2_N distribution at a significance value $\alpha = 0.05$. When $L = 1$, the D_1 spectrum is called a spectrum of string str deviation from homogeneity. At test-period λ of string str, the D_1 spectrum takes on the value

$$D_1(\lambda) = \frac{\psi(\Pi_{\text{str}}(\lambda), \Pi_{\text{Tdm}_1}(\lambda), n)}{\chi^2_{\text{crit}}((K-1)\lambda, \alpha)} \qquad (4)$$

where $\text{Tdm}_1 = \text{Tdm}_1(\Pi_{\text{str}}(1), n)$.

The hypothesis on the statistical indistinguishability of the strings str and $\text{Tdm}_L = \text{Tdm}_L(\Pi_{\text{str}}(L), n)$ is accepted if the condition $N_L/L_{\max} < 0.05$ is met, where $L_{\max} \sim n/5K$ and N_L is the number of test-periods at which the values of the spectrum $D_L > 1$. For example, as can be seen in Fig. 3, for the coding region of chicken *Gallus gallus* Apo A-IV mRNA, the hypothesis is accepted if $L = 33$, and it is rejected if $L = 3$.

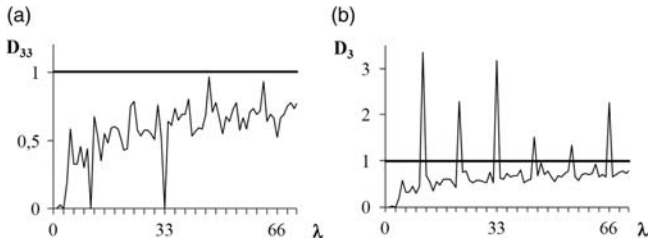

Figure 3. For the coding region of *G. gallus* Apo A-IV mRNA (GenBank Y16534, region: 37–1137 bp), the spectra of deviation from the proposed latent profility [Equation (3)]. (a) Deviation from the proposed 33-profility. (b) Deviation from the proposed 3-profility.

2.5. *Verification of periodicity pattern estimation*

To confirm the validity of the $\text{Str}_L(\Pi_{\text{str}}(L))$ estimation for the latent profility pattern in a textual string str, a reconstruction is built of the D_1 spectrum [Equation (4)] of string str deviation from homogeneity. The D_1 spectrum has been chosen as the most informative from the spectra pool of str deviation from the profility [Equation (3)].

The reconstruction is realized on the basis of the $\text{Str}_L(\Pi_{\text{str}}(L))$ pattern inducing the periodic profile string $\text{Tdm}_L = \text{Tdm}_L(\Pi_{\text{str}}(L), n)$. Thus, by analogy with Equation (4), for theoretical reconstruction of

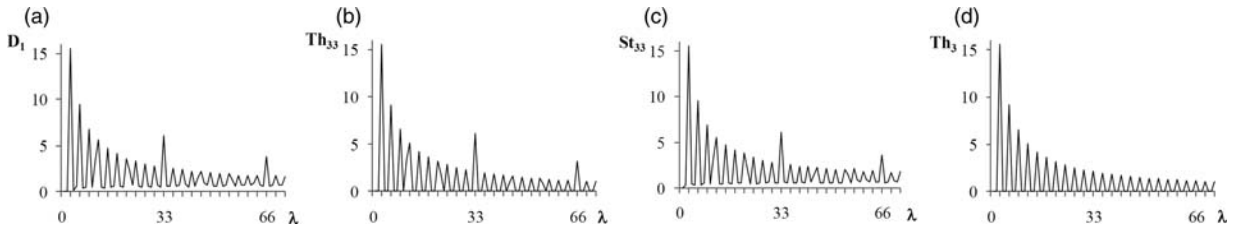

Figure 4. Verification of pattern estimation for latent profile periodicity of 33 bp (33-profility) in the coding region of *G. gallus* Apo A-IV mRNA (GenBank Y16534, region: 37–1137 bp). (a) Spectrum of deviation from homogeneity [Equation (4)]. (b) Theoretical [Equation (5)] and (c) statistical [Equation (6)] reconstructions of the spectrum in (a), under the assumption of 33-profility in the region. (d) Theoretical reconstruction of the spectrum in (a), under the assumption of 3-profility in the region.

the D_1 spectrum, the Th_L spectrum is chosen that at test-period λ of string str takes on the value

$$\mathrm{Th}_L(\lambda) = \frac{\psi(\Pi_{\mathrm{Tdm}_L}(\lambda), \Pi_{\mathrm{Tdm}_1}(\lambda), n)}{\chi^2_{\mathrm{crit}}((K-1)\lambda, \alpha)} \qquad (5)$$

If the Th_L spectrum follows the D_1 spectrum (Fig. 4a and b), then the $\mathrm{Str}_L(\Pi_{\mathrm{str}}(L))$ estimation of the pattern of latent L-profile periodicity (L-profility) in analysed textual string str is correct. Therefore, latent L-profile periodicity (L-profility) in string str is confirmed.

Instead of the theoretical reconstruction spectrum [Equation (5)], we can use the statistical (St_L) reconstruction of the D_1 spectrum (Fig. 4a and c). In this case, by using a random number generator and the main L-profile matrix $\Pi_{\mathrm{str}}(L)$ of the string Tdm_L, string str* is created as a realization of the string Tdm_L. Then, by analogy with Equation (4), the value of the St_L spectrum at the test-period λ is calculated as follows:

$$\mathrm{St}_L(\lambda) = \frac{\psi(\Pi_{\mathrm{str}*}(\lambda), \Pi_{\mathrm{Tdm}*_1}(\lambda), n)}{\chi^2_{\mathrm{crit}}((K-1)\lambda, \alpha)} \qquad (6)$$

where $\mathrm{Tdm}^*_1 = \mathrm{Tdm}_1(\Pi_{\mathrm{str}*}(1),\ n)$. Statistical reconstruction should be used when regular minima in the D_1 spectrum clear deviate from null.

3. Results and discussion

Methods of identifying a new type of latent periodicity in DNA called latent profile periodicity, or profility, have been proposed in the present work. A characteristic of this profile periodicity is the random nature of its pattern. The profile matrix of the pattern determines statistical periodicity in the appearance of the characters in textual strings. As a result, latent profile periodicity manifests in the analysed string.

3.1. *Profile-statistical basis of structural domains in protein families*

Application of the methods proposed in this work enabled us to discover the occurrence of latent profility for the 33 nucleotides (33-profility) in the coding gene regions of the PF01442 apolipoprotein family from the Pfam (http://pfam.sanger.ac.uk/) database of protein families. This family contains the apolipoproteins Apo A, Apo C and Apo E, which are members of a multigene family that probably evolved from a common ancestral gene. Apolipoproteins function in lipid transport as structural components of lipoprotein particles, cofactors for enzymes and ligands for cell-surface receptors. The family contains more than 800 protein sequences from ~100 species. In each position of the family, multiple alignment shows an average identity of amino acids of ~30%. By taking this apolipoprotein family as a case study, we can demonstrate a procedure for identifying latent profility.

The characteristic spectra of coding regions for the apolipoproteins Apo E of house mouse *Mus musculus*, Apo A-I of gilt-head sea bream *Sparus aurata* and Apo A-IV of chicken *G. gallus* are shown in Fig. 2a–c. In these spectra, the first clear maximum is found at the test-period of the 33 base pairs (bp). Thus, the estimation of a pattern size equal to the 33 bp is proposed. The maximum values in the spectra of deviation from the 33-profility [Equation (3)] do not exceed a figure of one ($D_{33} < 1$) that illustrated in Fig. 3a for chicken Apo A-IV. Using the result for the considered coding regions, estimates of patterns for the 33-profile periodicity may be proposed. These estimates are determined by a sample 33-profile matrix of the corresponding analysed region. The reasonableness of each pattern estimate is confirmed by similarity between the spectrum of deviation from homogeneity and its theoretical, or statistical, reconstruction for the analysed coding region. An example of verification of pattern estimation for latent 33-profility in the chicken Apo A-IV coding region is shown in Fig. 4. Comparison of Fig. 4a with d disproves the presence of latent 3-profility in the region, though a peak

at frequency 0.33 corresponding to the test-period of 3 bp dominates in the Fourier spectrum (Fig. 2d).

To check a robustness of the latent 33-profility pattern estimation found for the coding regions of apolipoproteins the damages in different consecutive segments of 33 bp have been simulated. Every kth ($k = 5,4,3,2$) segment of the coding region has been substituted by a fragment (of the same length) from homogeneous sequence with equally probable distribution of the nucleotides. The example of analysis of gene sequence with such a noise is shown in Fig. 5 for Apo A-IV mRNA of chicken *G. gallus*. Spectral–statistical analysis reveals a cutoff in pattern recognition when the sequence damages became equal to 25% ($k = 4$). In this case, there is a dilemma that what pattern size should be chosen—33 or 66 bp. According to the theoretical reconstructions (Fig. 5c and d) of the spectrum of deviation from homogeneity (Fig. 5b), an estimate of 66 bp appears to be more preferable. Such a preference becomes obvious with 50% of damages when the latent profile periodicity of 66 bp arises naturally. In the analysis done, no

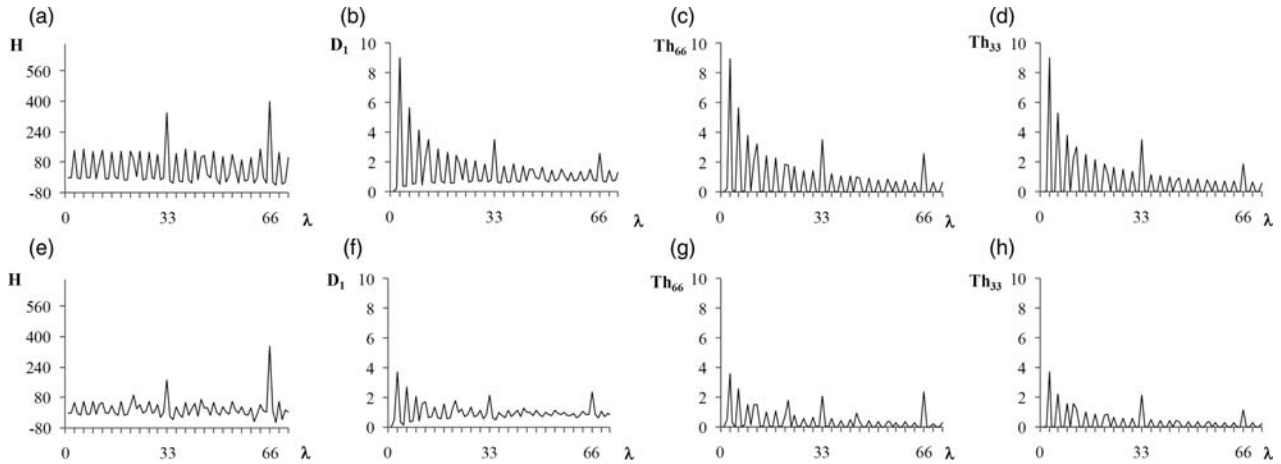

Figure 5. Robustness analysis of pattern estimation for the 33-profility in the coding region of *G. gallus* Apo A-IV mRNA (GenBank Y16534, region: 37–1137 bp). See Fig. 2c for the original characteristic spectrum of the region. Upper series: spectral-statistical analysis of the region sequence containing 25% of the destroyed 33-segments. Lower series: analysis of the region sequence containing 50% of the destroyed 33-segments. For the corresponding analysed sequences: (a and e) characteristic spectrum [Equation (2)], (b and f) the D1 spectrum [Equation (4)] of sequence deviation from homogeneity, (c and g) theoretical reconstruction [Equation (5)] of the D1 spectrum under supposition of the presence of 66-profility in the analysed sequence, (d and h) theoretical reconstruction [Equation (5)] of the D1 spectrum under supposition of the presence of 33-profility in the analysed sequence.

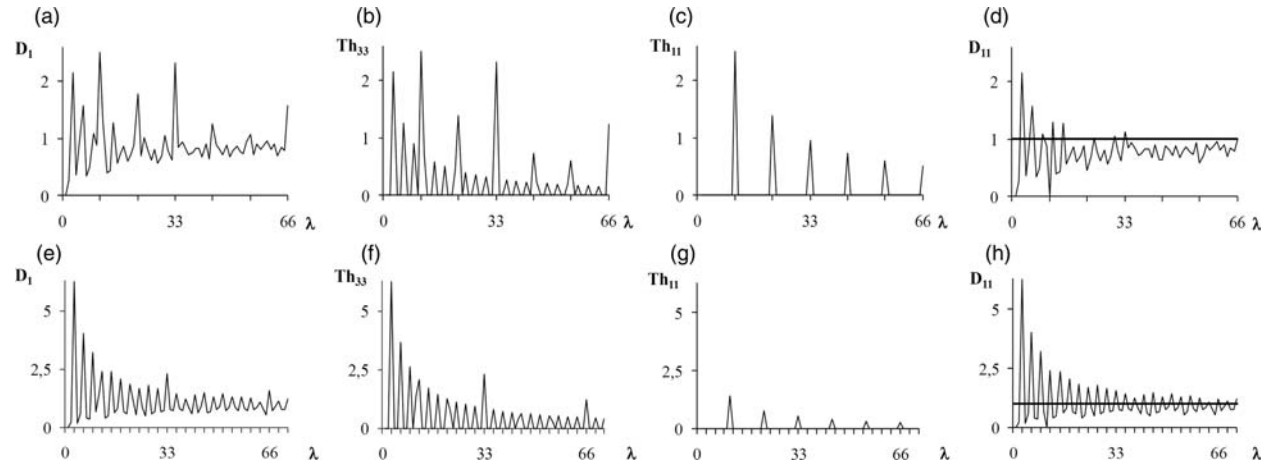

Figure 6. Upper series: verification of pattern estimation for latent profile periodicity of 33 bp (33-profility) in the central part of *M. musculus* Apo E mRNA (GenBank M12414, region: 275–604 bp). Lower series: verification of pattern estimation for latent 33-profility in the whole Apo E mRNA (GenBank M12414, region: 1–936 bp). For the corresponding analysed sequences: (a and e) the D1 spectrum [Equation (4)] of sequence deviation from homogeneity, (b and f) theoretical reconstruction [Equation (5)] of the D1 spectrum under supposition of the presence of 33-profility in an analysed sequence, (c and g) theoretical reconstruction [Equation (5)] of the D1 spectrum under supposition of the presence of 11-profility in an analysed sequence, (d and h) the D11 spectrum [Equation (4)] of sequence deviation from the 11-profility.

essential regions in the apolipoprotein gene sequences were revealed which determinate the occurrence of latent profility in the genes. The latent 33-profility of the apolipoprotein genes seems to be a consequence of consistent statistical low of their structural organization.

In earlier work,[23] after diagonal dot matrix analysis of internal homology within *M. musculus* Apo E mRNA, it was concluded that gene evolution took place by duplication of an 11-bp ancestral sequence. The supposition was also made that, before the genes of Apo E, Apo A-I and Apo A-IV were formed by duplication of the general 33-bp unit, copies of the ancient 11-'pattern' in the tandem 33-repeat underwent essential mutational alterations. In mouse Apo E mRNA fragment (GenBank M12414, 275–604 bp), 3-, 11- and 33-profility were examined in the present work. It is this fragment for which an ancestral 11-bp sequence was derived previously.[23] Using the methods proposed here, only 33-profility has been revealed and confirmed for the fragment. The same conclusion is made for the entire mouse Apo E mRNA sequence. Fig. 6 illustrates the performed analysis. Theoretical reconstructions of the D_1 spectra of deviation from homogeneity (Fig. 6a and e) were undertaken with the assumption of the occurrence of latent 33-profility in both the particular fragment and the entire Apo E mRNA (Fig. 6b and f, respectively), follow the corresponding D_1 spectra. Theoretical reconstructions (Fig. 6c and g) of the same D_1 spectra, undertaken with the assumption of the presence of latent 11-profility in the analysed sequences, are not similar to the corresponding D_1 spectra. Moreover, the D_{11} spectra of deviation from 11-profility (Fig. 6d and h) at numerous test-periods exceed the threshold ($D_{11} > 1$), indicating the absence of 11-profility in the analysed sequences. Domination of some nucleotides (more than 50%), revealed earlier[23] in four positions (the 2nd, 3rd, 7th and 9th) of the quasi-pattern of the 11 bp, is probably due to structural particularity of textual 33-repeat from the central part (275–604 bp) of mouse Apo E mRNA. In contrast, 33-profility undoubtedly settles a fixed periodicity of appearance for the nucleotides both in the central part and over the whole analysed Apo E mRNA.

The known secondary structure of the apolipoprotein family PF01442 contains several pairs of α helices of the 11 and 22 amino acid residues. Such a spatial organization correlates with the 33-bp profile periodicity of the apolipoprotein genes. The generic size of the pattern of latent profile periodicity in the PF01442 family genes possibly influences the formation of the typical secondary structure for the protein family and agrees well with the hypothesis on family origin from a common ancestral gene.

A generic size of ~290 bp for the latent profile periodicity pattern is observed in the genes of fibronectin type III domain-containing protein, which is a protein of an intercellular matrix. It is a glycoprotein that many cells synthesize and secrete into intercellular space. The fibronectin consists of two identical polypeptide chains joined by disulfide bridges near the C-terminuses. Each polypeptide chain contains ~10 domains, each of which holds the specific sites binding the various substances. The proposed spatial structure of the domain contains seven antiparallel β-strands.[24]

Orthologous genes of the fibronectin type III domain family that have been analysed in the present work are listed in Table 1. Table 2, using data from the KEGG database (http://www.genome.jp/kegg/), shows an identity percentage between pairs of the protein family. In characteristic spectra (e.g. see Fig. 7a) of genes from this family, the generic pattern size of latent profile periodicity is found to be 291 bp. Statistical reconstruction (Fig. 7c) corresponding to the pattern size of 291 bp reconstitutes the spectrum of deviation from homogeneity (Fig. 7b), which verifies latent profile periodicity with this period. The pattern size of 291 bp is in good agreement with the size of repeated domains (~90–100 amino acids) in the proteins of the family.

Table 1. Analysed orthologous genes of the fibronectin type III domain family

Organism	KEGG entry	CDS (bp)	Protein	Number of domains
Homo sapience (human)	hsa:22862	3597	Fibronectin type III domain containing protein 3A, 1198 amino acids	9
Mus musculus (mouse)	mmu:319448	3597	Fibronectin type III domain containing protein 3A, 1198 amino acids	9
Gallus gallus (chicken)	gga:418863	3600	Fibronectin type III domain containing protein 3A, 1199 amino acids	9
Xenopus laevis (frog)	xla:446899	3600	Fibronectin type III domain containing protein 3A, 1199 amino acids	9

Table 2. Identity percentage between the protein pairs from Table 1 according to the KEGG database

	hsa:22862	mmu:319448	gga:418863	xla:446899
hsa:22862	100.0	91.1	80.2	53.3
mmu:319448	91.1	100.0	68.5	52.3
gga:418863	80.2	68.5	100.0	53.1
xla:446899	53.3	52.3	53.1	100.0

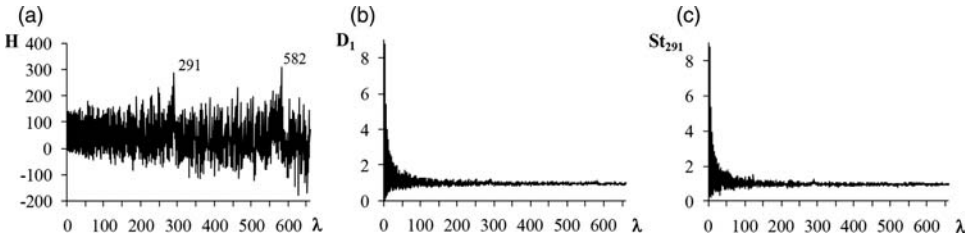

Figure 7. Identification of latent profility of 291 bp in the gene coding region of fibronectin type III domain-containing protein (KEGG GENES hsa:22862). (a) The characteristic spectrum. (b) The D1 spectrum of deviation from homogeneity (1-profility). (c) The statistical reconstruction of the D1 spectrum carried out assuming the presence of 291-profility in the sequence.

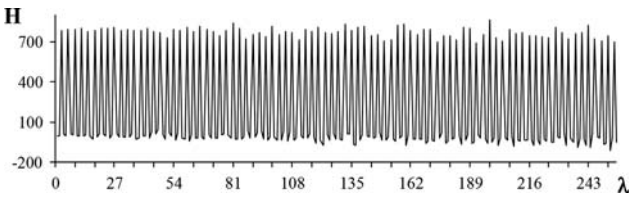

Figure 8. The characteristic spectrum for the coding region of *cya* gene from bacterium *B. pertussis* (GenBank Y00545, region: 981–6101 bp).

3.2. *Manifestation of levels of organization of genetic information encoding*

Regularity of the peaks at 3 bp is observed in the characteristic spectra of the coding regions (Figs 2a–c, 7a and 8). Thus, an encoding organization level caused by the genetic triplet code is manifested. This regularity of a characteristic spectrum is called further as 3-regular heterogeneity, or 3-regularity. As in the Fourier spectra, 3-regular heterogeneity of a characteristic spectrum can be observed in the absence of latent periodicity of 3 bp (see Fig. 2, for example). If 3-regularity exists in the characteristic spectrum of a coding region, then revealing latent profility (different from 3-profility) in the spectrum determines the second level of the encoding organization. For example, Fig. 2a–c shows the characteristic spectra in which, as discussed above (Figs 3 and 4), latent 33-profility is revealed against the background of 3-regularity.

An investigation of different levels in the organization of genetic information encoding has been carried out on a sample of 18 140 human coding regions (CDS) from the KEGG GENES-54.1 database (http://www.genome.jp/kegg/genes.html). Only those coding regions were chosen for which there is experimental evidence of protein translation. Open reading frames, hypothetical proteins, tRNA and rRNA, and genes assumed by their sequences to show similarity to other known genes were excluded from the sample. It appears that 3-regularity in the characteristic spectra is fixed for 93% of the sample (16 786 CDS). Against the background of 3-regular heterogeneity, latent 3-profility is revealed for 62% (11 200 CDS) of the original sample. For the 11% of the sample (1953 CDS), two levels of organization of the encoding are manifested (3-regular heterogeneity and the latent profility different from 3-profility).

Taking into account the inaccuracy of the statistical methods, the following conclusions can be made from the results obtained. Owing to amino acids triplet encoding, the 3-regular heterogeneity of the characteristic spectra is generic for human genes. However, such regularity is not due to latent periodicity of 3 bp. Thus, it is essential to differentiate between the phenomena of regular heterogeneity and latent periodicity in the genetic sequences. In order to verify the existence of latent periodicity of some type, it is necessary to observe a pattern inducing the periodicity.

3.3. *Local profile periodicity*

In the coding regions, the manifestation of tha local two-level organization of genetic information encoding is possible. Thus, for the whole coding region, 3-regular heterogeneity only is observed (see Fig. 8, for example). Regions with local profile periodicity (local profility) can be revealed by scanning the sequence with a small window. For example, regarding local profility in the coding region of the *cya* gene

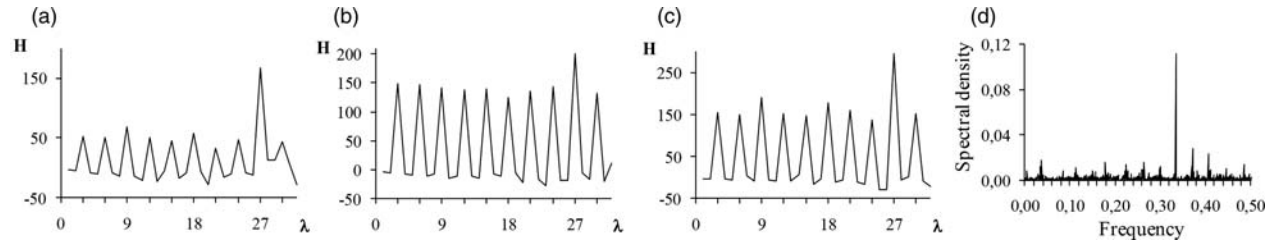

Figure 9. Characteristic spectra for the three local areas of the coding region of *cya* gene from bacterium *B. pertussis* (GenBank Y00545, region: 981–6101 bp). (a) Local area 4020–4181 bp. (b) Local area 4443–5036 bp. (c) Local area 5211–5840 bp. (d) Fourier spectrum for the local area (c).

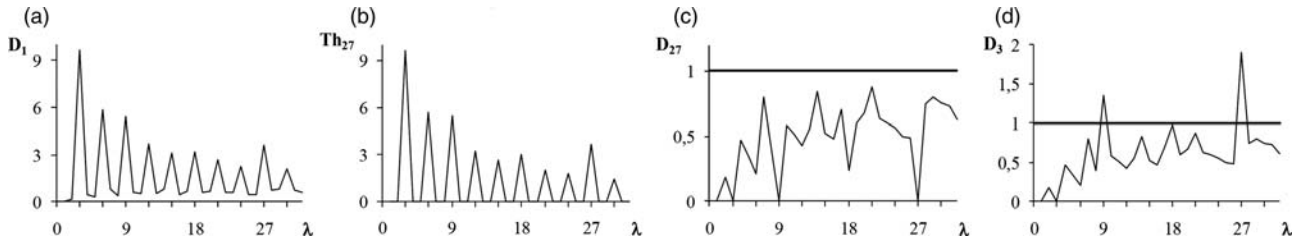

Figure 10. Identification of latent 27-profility in a local area of the coding region of the *cya* gene from bacterium *B. pertussis* (GenBank Y00545, 5211–5840 bp). (a, c and d) Spectra of deviation from the 1-, 27- and 3-profility (respectively). (b) A spectrum of theoretical reconstruction of the D1 spectrum in (a) assuming the presence of 27-profility in the local area.

from the bacterium *Bordetella pertussis* (GenBank Y00545, 981–6101 bp), in the entire coding region, 3-regular heterogeneity only is revealed (Fig. 8), and there is no latent profility. Latent profile periodicity is observed solely in local areas of the coding regions. Three local areas of latent profile periodicity with a period of 27 bp (Figs 9a–c and 10) can be distinguished in the coding region of the *cya* gene of bacterium *B. pertussis*.

Let us note again that the first level of encoding is manifested in the 3-regularity of characteristic spectrum (Fig. 9a–c) and in existence of dominant peak at frequency equal to 0.33 in the Fourier spectrum (see, for example, Fig. 9d). The second level of encoding organization—the 27-profile periodicity—in the local areas of the *cya* gene is pointed at by the dominant peaks of characteristic spectra. Such a profile periodicity is proved by reconstruction of the spectrum of deviation from homogeneity in every local area (see, for example, Fig. 10a and b). In contrast to the characteristic spectra, the second level of encoding organization is not manifested in the Fourier spectra (see, for example, Fig. 9d).

The *cya* gene encodes bifunctional hemolysin/adenylate cyclase (UniProtKB P15318) in which the areas corresponding to the gene local 27-profility hold the hemolysin-type calcium-binding sites. These sites have a periodic structure of 18 amino acid residues (Fig. 11) corresponding to 54 bp (2 × 27 bp) in the gene.

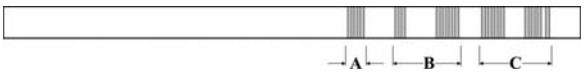

Figure 11. Schematic representation of bifunctional hemolysin/adenylate cyclase (UniProtKB P15318, 1–1706 amino acids) encoded by the cya gene of bacterium *B. pertussis* (GenBank Y00545, region: 981–6101 bp). Each grey vertical bar denotes a hemolysin-type calcium-binding site of 18 amino acids. The regions A (1014–1067 amino acids), B (1155–1352 amino acids) and C (1411–1620 amino acids) correspond to the three areas of the local 27-profility (4020–4181, 4443–5036 and 5211–5840 bp, respectively) revealed in the *cya* gene (Fig. 9).

3.4. Conclusions

Methods for identifying a new type of latent periodicity—latent profility in DNA—have been proposed. For DNA coding regions, latent profility enables us to distinguish two levels of organization of genetic information encoding. The first level (the triplet level of encoding), revealed via the Fourier analysis techniques, indicates the phenomenon of regular heterogeneity in the DNA coding regions. The second level of organization in the encoding is due to latent profile periodicity of the DNA sequence. It has been shown that latent profile periodicity in genes of the same family may correlate with the structural features of encoded proteins. Such an effect may manifest in the local areas of coding regions where latent profile periodicity is observed.

References

1. Lobzin, V.V. and Chechetkin, V.R. 2000, Order and correlations in genomic DNA sequences. The spectral approach, *Physics–Uspekhi.*, **43**, 55–78.
2. Li, W. 1997, The study of correlation structures of DNA sequences: a critical review, *Comput. Chem.*, **21**, 257–71.
3. Shepherd, J.C.W. 1981, Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification, *Proc. Natl Acad. Sci. USA*, **78**, 1596–600.
4. Benson, G. 1999, Tandem repeats finder: a program to analyze DNA sequences, *Nucleic Acids Res.*, **27**, 573–80.
5. Metzgar, D., Bytof, J. and Wills, C. 2000, Selection against frameshift mutations limits microsatellite expansion in coding DNA, *Genome Res.*, **10**, 72–80.
6. Borstnik, B. and Pumpernik, D. 2002, Tandem repeats in protein coding regions of primate genes, *Genome Res.*, **12**, 909–15.
7. Trifonov, E.V. 1987, Translation framing code and frame-monitoring mechanism as suggested by the analysis of mRNA and 16S RNA nucleotide sequences, *J. Mol. Biol.*, **194**, 643–52.
8. Tsonis, A.A., Elsner, J.B. and Tsonis, P.A. 1991, Periodicity in DNA coding sequences: implications in gene evolutions, *J. Theor. Biol.*, **151**, 323–31.
9. Gutierrez, G., Oliver, J. and Marin, A. 1994, On the origin of the periodicity of three in protein coding DNA sequences, *J. Theor. Biol.*, **167**, 413–14.
10. Fickett, J.W. and Tung, C.-S. 1992, Assessment of protein coding measures, *Nucleic Acid Res.*, **20**, 6441–50.
11. Silverman, B.D. and Linsker, R. 1986, A measure of DNA periodicity, *J. Theor. Biol.*, **118**, 295–300.
12. Tiwari, S., Ramachandran, S., Bhattacharya, A., Bhattacharya, S. and Ramaswamy, R. 1997, Prediction of probable genes by Fourier analysis of genomic sequences, *Comput. Appl. Biosci.*, **13**, 263–70.
13. Issac, B., Singh, H., Kaur, H. and Raghava, G.P.S. 2002, Locating probable genes using Fourier transform approach, *Bioinformatics*, **18**, 196–97.
14. Chaley, M.B. and Kutyrkin, V.A. 2008, Model of perfect tandem repeat with random pattern and empirical homogeneity testing poly-criteria for latent periodicity revelation in biological sequences, *Math. Biosci.*, **211**, 186–204.
15. Korotkov, E.V., Korotkova, M.A. and Kudryashov, N.A. 2003, Information decomposition method to analyze symbolical sequences, *Phys. Lett. A*, **312**, 198–210.
16. Gatherer, D. and McEwan, N.R. 2003, Analysis of sequence periodicity in *E. coli* proteins: empirical investigation of the "duplication and divergence" theory of protein evolution, *J. Mol. Evol.*, **57**, 149–58.
17. Larsabal, E. and Danchin, A. 2005, Genomes are covered with ubiquitous 11 bp periodic patterns, the class A flexible patterns, *BMC Bioinformatics*, **6**:206.
18. Chaley, M.B. and Kutyrkin, V.A. 2010, Structure of proteins and latent periodicity in their genes, *Moscow Univ. Biol. Sci. Bull.*, **65**, 133–35.
19. Kolpakov, R. and Kucherov, G. 2003, mreps: efficient and flexible detection of tandem repeats in DNA, *Nucleic Acids Res.*, **31**, 3672–78.
20. Parisi, V., Fonzo, V.D. and Aluffi-Pentini, F. 2003, STRING: finding tandem repeats in DNA sequences, *Bioinformatics*, **19**, 1733–38.
21. Mudunuri, S.B. and Nagarajaram, H.A. 2007, IMEx: imperfect microsatellite extractor, *Bioinformatics*, **23**, 1181–87.
22. Sharma, D., Issac, B., Raghava, G.P.S. and Ramaswamy, R. 2004, Spectral Repeat Finder (SRF): identification of repetitive sequences using Fourier transformation, *Bioinformatics*, **20**, 1404–12.
23. Rajavashisth, T.B., Kaprein, J.S., Reue, K.L. and Lusis, A.J. 1985, Evolution of apolipoprotein E: mouse sequence and evidence for an 11-nucleotide ancestral unit, *Proc. Natl Acad. Sci. USA*, **82**, 8085–89.
24. Bazan, J.F. 1990, Structural design and molecular evolution of a cytokine receptor superfamily, *Proc. Natl Acad. Sci. USA*, **87**, 6934–38.