Chapter 16

Spectral–Statistical Approach for Revealing Latent Regular Structures in DNA Sequence

Maria Chaley and Vladimir Kutyrkin

Abstract

Methods of the spectral–statistical approach (2S-approach) for revealing latent periodicity in DNA sequences are described. The results of data analysis in the HeteroGenome database which collects the sequences similar to approximate tandem repeats in the genomes of model organisms are adduced. In consequence of further developing of the spectral–statistical approach, the techniques for recognizing latent profile periodicity are considered. These techniques are basing on extension of the notion of approximate tandem repeat. Examples of correlation of latent profile periodicity revealed in the CDSs with structural–functional properties in the proteins are given.

Key words Latent periodicity, Approximate tandem repeats, Profile periodicity, HeteroGenome database, CDS, Spectral–statistical approach

1 Introduction

Until recently the reliable methods for recognizing latent periodicity in genome were based on the notion of approximate tandem repeat [1, 2]. However, employment of these methods has shown that approximate tandem repeats constitute a small part in the genome sequences of various organisms. So, the indirect methods for estimating latent periodicity period have spread, exploited without determination of periodicity type and its corresponding pattern. Fourier analysis [3–7] and the other techniques [8–15] displaying dominant peaks in the graphs of a single statistical parameter which values depend on the tested periods of DNA sequence can be referred to such methods. Without a model of periodicity, the latent period estimate obtained by such methods cannot be unambiguously interpreted.

Spectral–statistical approach to revealing latent periodicity has been originally developed in the work [12]. Initially the problem was set to select quantitative statistical parameters for revealing approximate tandem repeats and DNA sequences that are similar

Oliviero Carugo and Frank Eisenhaber (eds.), *Data Mining Techniques for the Life Sciences*, Methods in Molecular Biology, vol. 1415, DOI 10.1007/978-1-4939-3572-7_16, © Springer Science+Business Media New York 2016

with the repeats. In investigating approximate tandem repeats in the TRDB database [16], two characteristic statistical parameters have been revealed. One of them characterized heterogeneity level that in approximate tandem repeats has sufficiently high values. Another one described a mean level of character (base) preservation at tested period. This mean level is close to unity (~0.8), if a tested period coincides with latent period in the approximate tandem repeats. In the framework of spectral–statistical approach (the 2S-approach), these statistical parameters are considered in accordance with a length of tested period in analyzed DNA sequence. The graphics of these parameters are called spectra. They characterize initial stage in the developing of the 2S-approach with methodology represented in the works [12, 17, 18].

The analysis of genome sequences from the model organisms *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, *Caenorhabditis elegans*, and *Drosophila melanogaster* has been done with the help of the 2S-approach spectra. In the result of the analysis, the HeteroGenome database (http://www.jcbi.ru/lp_baze/) has been created [18] for which the sequences similar to the approximate tandem repeats were selected from DNA sequences of the organisms. The description of the HeteroGenome database methodology will be done in the next sections.

However, according to the data from the HeteroGenome, DNA sequences similar to approximate tandem repeats cover a small part of genome (~10 %). So, the methods, searching for latent periodicity of unknown type, are widely spread that could be called indirect, as they are not based on any model of periodicity. For example, Fourier analysis and the like techniques can be placed to such methods [3–9]. Dominant peaks revealed by these methods in the spectra are used to estimate period length of latent periodicity. In the strict sense, such estimates of period length demand an additional instantiation [19].

A new notion of latent periodicity called latent profile periodicity has been proposed in the works [12, 20]. This new notion is based on a model of profile periodicity (profility) [20, 21] allowing generalize notion of approximate tandem repeat. Basing on this model, the 2S-approach has got a shot in the arm of recognizing the latent profile periodicity in DNA sequences [21, 22]. Since new type of periodicity generalizes the notion of approximate tandem repeat, one can suppose a share of recognizable latent periodicity will sufficiently grow. This assumption is proved by the examples of analysis of DNA sequences from human genome [21]. The results of the analysis allowed putting forward a hypothesis about the existence of two-level organization of encoding in the CDSs. Besides, it appears that latent profility, revealed in coding DNA regions, can be translated into structural particularities of protein sequence. Direct revelation of such particularities is a sufficiently complicated problem because the goal of the search is a priori unknown.

New methods of the 2S-approach have been proposed [20–22] for recognizing latent profile periodicity. They are based on a model of profile string that is special periodic random string with a pattern of independent random characters. Every one of such the random characters is a random variable taking on the values from textual alphabet of DNA sequences. In the frames of the 2S-approach, DNA sequence with displayed latent profile periodicity is considered as realization of a profile string. Therefore, statistical methods and criteria have to be used for recognizing latent profile periodicity. Existence of latent profile periodicity in DNA sequence is recognized in that case, when this sequence is statistically close to a profile string. In fact, the problem of latent profile periodicity recognition in DNA sequence leads to the problem of specifying a profile string considered as periodicity etalon for the sequence. Random pattern of such a profile string is an analogue of consensus-pattern deduced from the sequence of approximate tandem repeat. One of the next sections is deduced to the description of the 2S-approach for recognizing latent profile periodicity.

2 HeteroGenome Database. Materials, Methodology, and Analysis of the Results

The methods of the 2S-approach to search for the regions in DNA sequence that are close to approximate tandem repeats have been applied to the genome sequences of well-studied model organisms [23] *S. cerevisiae*, *A. thaliana*, *C. elegans*, and *D. melanogaster*. These organisms represent a genome of the eukaryotes ranging from unicellular organism (baker's yeast) to multicellular plants (*Arabidopsis*) and animals (nematode), which facilitates the general study of the phenomenon of latent periodicity in genome. Original DNA sequences of the whole genomes of model organisms have been obtained from the GenBank [24] at ftp://ftp.ncbi.nih.gov/genomes/. The results of genome analysis have been systemized in the HeteroGenome database (http://www.jcbi.ru/lp_baze/) described in the work [18].

Approximate tandem repeats are the most studied type of latent periodicity in DNA sequences, because this type is described by relevant models [1, 2]. A significant number of publications are devoted to search for approximate tandem repeats and their recognition (e.g., *see* Refs. [25–28]). However, such repeats constitute sufficiently small part in genome sequences of various organisms [18]. Besides, the methods, estimating length of latent period in the sequences which are not approximate tandem repeats, gained widespread acceptance in scientific literature (e.g., *see* Ref. [9]). At that, type of periodicity remains unknown, and it is not based on any model. So, in creating the HeteroGenome database, the following compromise approach to search for the sequences with latent periodicity was chosen. The sequences similar to approximate tandem repeats were selected. As similarity estimate two parameters have been chosen whose high values are characteristic for the periods of approximate tandem repeats. These parameters will be further described in detail.

2.1 Spectral–Statistical Approach for Revealing DNA Sequences Similar to Approximate Tandem Repeats

The revelation of latent periodicity close to approximate tandem repeats was done by determining heterogeneity of high significance level $\left(\sim 10^{-6}\right)$ at the test periods of an analyzed nucleotide sequence. A test period of DNA sequence is called an integer number which does not exceed one-half the sequence length. For each test-period λ analyzed sequence is divided into the substrings of length λ (last substring can be of smaller length).

Division into the substrings of length λ allows calculating a frequency $\pi_j^i \le 1 \left(i = \overline{1,4},\ \ j = \overline{1,\lambda}\right)$ to find a character a_i from nucleotide sequence alphabet $A < a = a_1,\ t = a_2,\ g = a_3,\ c = a_4 >$ in the jth position of the test period λ. Matrix $\pi = \left(\pi_j^i\right)_\lambda^K$ is called a sample λ-profile matrix for analyzed sequence, where $K = 4$ is the size of alphabet A. Then in analyzed sequence *a character preservation level* $pl(\lambda)$ at the test-period λ is determined by a formula:

$$pl\left(\lambda\right) = \frac{1}{\lambda}\sum_{j=1}^{\lambda} \max\left\{\pi_j^i :\ i \in 1,\ldots,K\right\}. \qquad (1)$$

By such a way, for an analyzed sequence at its test periods, a spectrum of character preservation level **pl** is introduced. According to the results of numerical experiments [12], character preservation level $pl\left(L\right) \ge 0.5$ corresponds to the sequences of approximate tandem repeats with period length equal to L.

Along with the high value of the **pl** spectrum, high level of the repeat's heterogeneity is observed at period length in approximate tandem repeat. In the HeteroGenome, a check on heterogeneity in the sequence of length n at the test-period λ is done with the help of Pearson χ^2-statistics [29]:

$$\nu\ \left(\lambda,n\right) = \frac{n}{\lambda}\sum_{j=1}^{\lambda}\sum_{i=1}^{K}\left(\pi_j^i - p^i\right)^2 / p^i\left(1 - p^i\right). \qquad (2)$$

In accordance with the results of numerical experiments done in the work [12], high character preservation level allows omitting claim of a large number of the repeats for the test-period λ. When character preservation level is high ($pl\left(\lambda\right) \sim 0.8$ and more), a value of the statistics (Eq. 2) is not taken into consideration, even though the number of repeats $\frac{n}{\lambda} < 5$ is small.

In searching for the sequences similar to approximate tandem repeats, check on heterogeneity in DNA sequence is carried out at a level of significance $\alpha = 10^{-6}$ [12]. For the test-period L, a critical value $\chi_{crit}^2(\alpha, N)$ with $N = \left(K - 1\right)\left(L - 1\right)$ freedom degrees corresponds to this level. If character preservation level $pl(L)$ is sufficiently high and value of statistics $\nu(L, n)$ meets a condition

$$v(L,n) / \chi^2_{crit}(\alpha,(K-1)(L-1)) \geq 1, \tag{3}$$

then the sequence is recognized similar to approximate tandem repeat with period *L*. In this case it is supposed that the value of *pl*(*L*) is close to a maximal value of the **pl**-spectrum in a range of the test periods of the sequence. So, as spectral characteristics of analyzed nucleotide sequence in the HeteroGenome database, a spectrum **H** is used that at the test-period λ takes on a value

$$H(\lambda) = v(\lambda,n) / \chi^2_{crit}(\alpha,(K-1)(\lambda-1)), \ \alpha = 10^{-6}. \tag{4}$$

The graphic of the **H**-spectrum obviously demonstrates a display of significant heterogeneities in a sequence at those test periods, where $H(\lambda) > 1$, and these test periods are further analyzed with the help of the **pl**-spectrum. As it was mentioned above, one of these test periods is selected as an estimate for the period length of latent periodicity that is pointed at by the first clear-cut maximal value in the **pl**-spectrum (*see* Fig. 1). Such a maximal value of the **pl**-spectrum can be interpreted as an index of preservation for the copies of periodicity pattern. Figure 1 gives an example of how, by jointly using both of the parameters (**H**-spectrum and

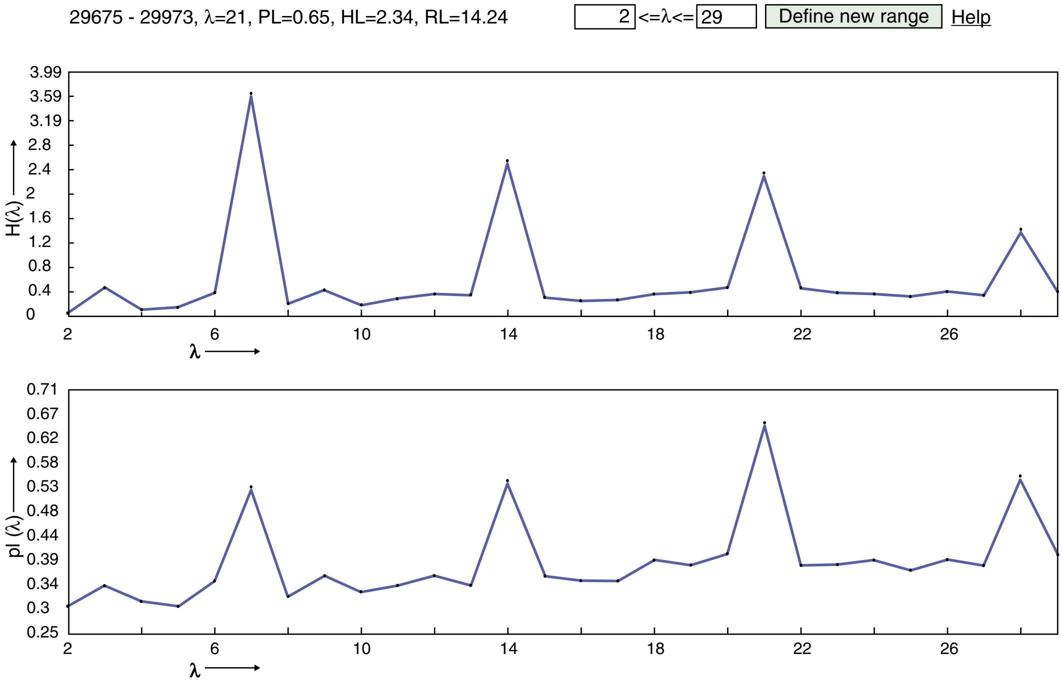
29675 - 29973, λ=21, PL=0.65, HL=2.34, RL=14.24

Fig. 1 The spectral–statistical characteristics in the HeteroGenome database for DNA sequence from *C. elegans* chromosome V (29675–29973 bps). At the *top*: spectrum of heterogeneity display (**H**-spectrum, *see* Eq. 4). At the *bottom*: spectrum of character preservation level (**pl**-spectrum, *see* Eq. 1). Maximal peak at 21 bp in the **pl**-spectrum corresponds to period length of the latent periodicity

pl-spectrum), one can unambiguously estimate periodicity pattern length. The analysis of a graphic of the **H**-spectrum in Fig. 1 allows distinguishing heterogeneities in a sequence under consideration at the test-periods multiple of seven. Maximal value in the **pl**-spectrum outlines the test period of 21 bp which is accepted as an estimate of periodicity pattern length. So, in the HeteroGenome database, visualization of the sequence alignment at the test period of 21 bp is shown automatically. User can additionally obtain the sequence alignment at the other test periods.

2.2 Strategy of Searching for and Structuring Data in the HeteroGenome

In creating the HeteroGenome database [18], to reveal periodicity close to approximate tandem repeats, a method of searching for DNA regions with highly significant heterogeneity (at the level $\alpha = 10^{-6}$), by scanning a series of overlapping windows, has been applied. Length of initial window is equal to 30 bp. Length of each the following window is set twice as large, until a limiting value will be achieved. Shifting with variable step, the windows scan an analyzed DNA sequence. General strategy of searching for the sequences similar to approximate tandem repeats resembled "shotgun strategy" of genome sequencing [30]. Within the framework of such a strategy, relatively short and overlapping fragments are sequenced first. Then computer assembling of the fragments into the more extended regions is done, and the borders of revealed heterogeneity regions are optimized.

For nonredundant data representation in the HeteroGenome database, each logical record is a group of DNA sequences revealed on chromosome with statistically significant heterogeneity (latent periodicity) which are intersected or (and) have the same or multiple period length. There are two levels of data representation in the group. At the first level, DNA sequence of the greatest length is considered that is called group representative. The rest sequences belong to the second level. As a rule, they correspond to the well-determined local structures of periodicity in the sequence of group representative.

2.3 Results of the HeteroGenome Data Analysis

The comparison of the data on periodicity for the genomes of *S. cerevisiae*, *A. thaliana*, *C. elegans*, and *D. melanogaster* in the HeteroGenome with corresponding data in the TRDB database [16] has shown that the HeteroGenome collects practically all tandem repeats represented in the TRDB and, moreover, essentially supplements them with the data on highly divergent tandem repeats.

In investigating the evolution and functional meaning of the latent periodicity regions in genome, the proportion of the whole genome covered by such regions is a quantitative indicator of no little significance. Nonredundant data on the regions of significant heterogeneity (latent periodicity) in the HeteroGenome database approximate tandem repeats (period length is of order 1000 bp), the latent periodicity regions in human genome account for about 10 % [25]. Also, taking into consideration data from the Table 1, it can be supposed that periodicity in eukaryotic genome constitutes ~10 %. Probably, such a percent is due to a balance between the molecular mechanism of originating tandem repeats and divergence of their sequences which stabilizes length of the repeats.

2.3.1 Impact of Latent Periodicity on Chromosome Length

Periodicity regions are the hot spots in genome, able to both expand and diminish size in response to slippage of DNA replicase and recombination and duplication processes [31–33]. Mutations (point substitutions, insertions/deletions of the nucleotides) disturb with time determined structure of DNA periodicity regions, stabilizing region lengths. Since the method of latent periodicity revelation used in the work [18] allows nonredundant estimating the periodicity proportion in genome, it becomes possible to investigate an influence of periodicity regions at the chromosomes.

Let us consider a percentage of periodicity regions in accordance with chromosome length in the genomes of analyzed model organisms (*see* Fig. 2). For each organism a characteristic scatter of the percents of chromosome's coverage by periodicity regions is observed. Though in the genomes of *S. cerevisiae*, *C. elegans*, and *D. melanogaster* a scatter of the percents for the chromosomes is comparable to a mean percent value in corresponding genome, in *A. thaliana* genome such a scatter is no more than 0.75 %. As Fig. 2a shows, while chromosome length is growing, the percent of the periodicity regions remains practically constant for *Arabidopsis* chromosomes.

Generally, as shown in Fig. 2, with growth of chromosome length, a percentage of its periodicity regions has a tendency to constancy or even reduction in all analyzed genomes of the model organisms. Nevertheless, in the consequence of ability for elongation, tandem repeats have markedly influenced at chromosome length (periodicity coverage ~10 %).

Table 1
Proportion of latent periodicity regions in the genomes of model organisms

Species	Genome length, bp	Total length of latent periodicity regions, bp	Percent of latent periodicity regions in genome, %
S. cerevisiae	12070900	419909	3.5
A. thaliana	119146348	4247672	3.6
C. elegans	100269917	6692629	6.7
D. melanogaster	120381546	5108483	4.2

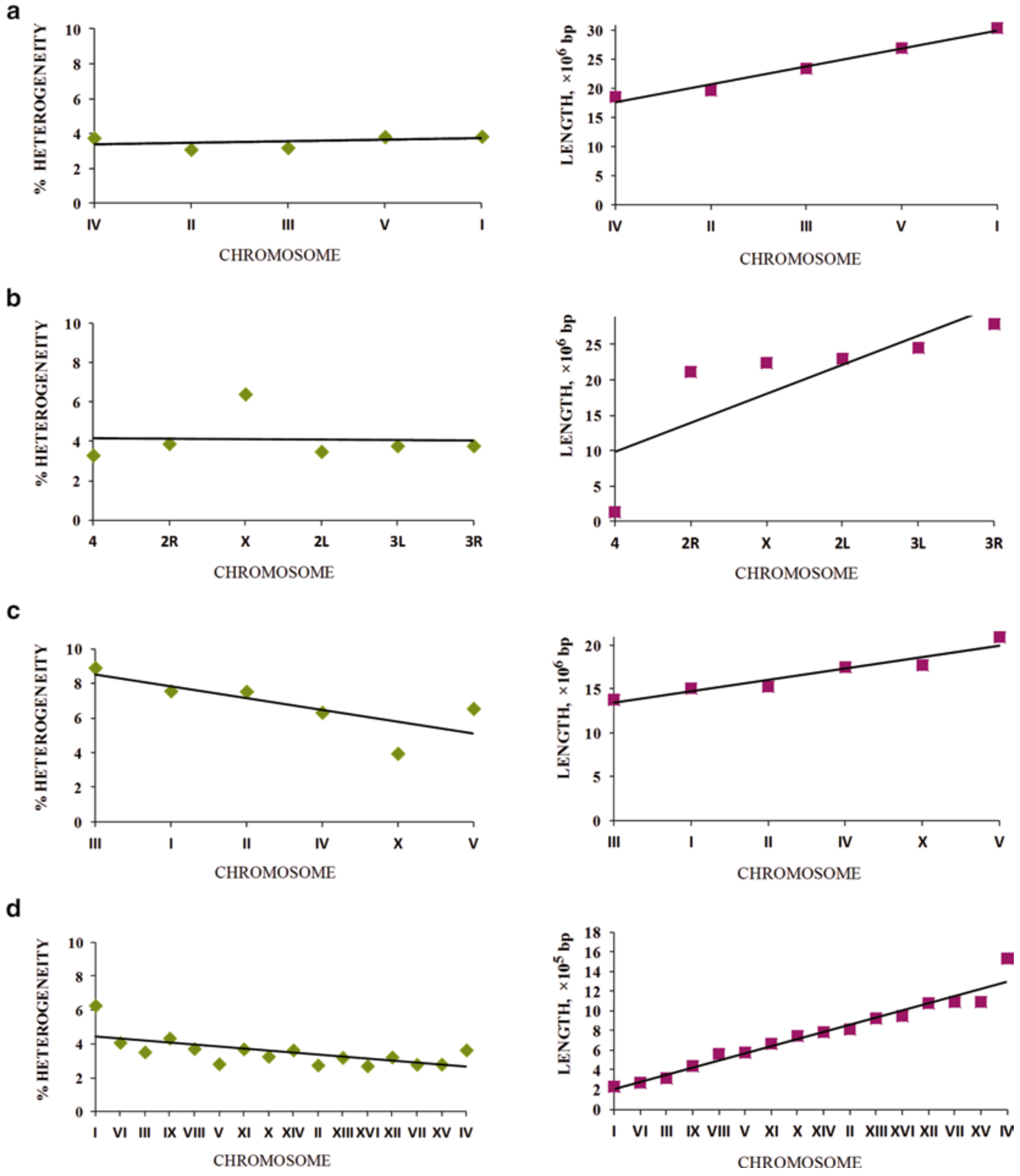

Fig. 2 Percentage of the latent periodicity (heterogeneity) regions on the chromosomes of model organisms of *A. thaliana* (**a**), *D. melanogaster* (**b**), *C. elegans* (**c**), and *S. cerevisiae* (**d**). The chromosomes of each organism are ordered by increase of their length, as shown in the *graphics on the right*. *Solid straight line* in the graphics designates a trend

allows estimating the percent of tandem repeats in the analyzed genomes of model organisms. Table 1 represents such estimates.

As it will be shown further, the largest part of latent periodicity regions in the analyzed genomes is represented by micro- and mini-satellites (period length is less than 100 bp). It is known that in human genome its fraction amounts to 3 % [30]. With the other

2.3.2 Analysis of Periodic Structure Preservation in the Regions of Heterogeneity

In accordance to the HeteroGenome data, Fig. 3 gives an example of histogram showing a distribution of the revealed latent periodicity regions in relation to preservation level of their periodic structure (*see* Eq. 1 for $pl(L)$ parameter). Separately for micro- (period length is in a range $2 \le L \le 10$), mini- $(10 < L \le 100)$, and mega- $(100 < L \le 2000)$ satellites for each chromosome, a percent of the repeats' length is shown for highly divergent $(0.4 \le pl \le 0.7)$, moderately $(0.7 \le pl \le 0.8)$, slightly $(0.8 < pl \le 0.9)$ divergent, and perfect $(0.9 < pl \le 1.0)$ tandem repeats.

According to Fig. 3, in the genome of *A. thaliana*, highly divergent mini-satellites $(10 < L \le 100)$ constitute a noticeable part $(\sim 1 - 1.5\ \%$ for each chromosome) which is comparable with the percentage of micro-satellites $(2 \le L \le 10)$. Consequently, mini- and micro-satellites similarly contribute into structural and functional organization of *A. thaliana* genome. A portion of mega-satellite repeats in *Arabidopsis* genome $(\sim 1\%)$ is also sufficiently noticeable.

On the page Database Statistics (http://www.jcbi.ru/lp_baze/statistics/index.html) in the HeteroGenome database, one can see analogous histograms for structural content of periodicity regions on the other chromosomes of the rest analyzed genomes. Basing on the analysis of these histograms, in every genome one or few types of characteristic dominating periodicities can be

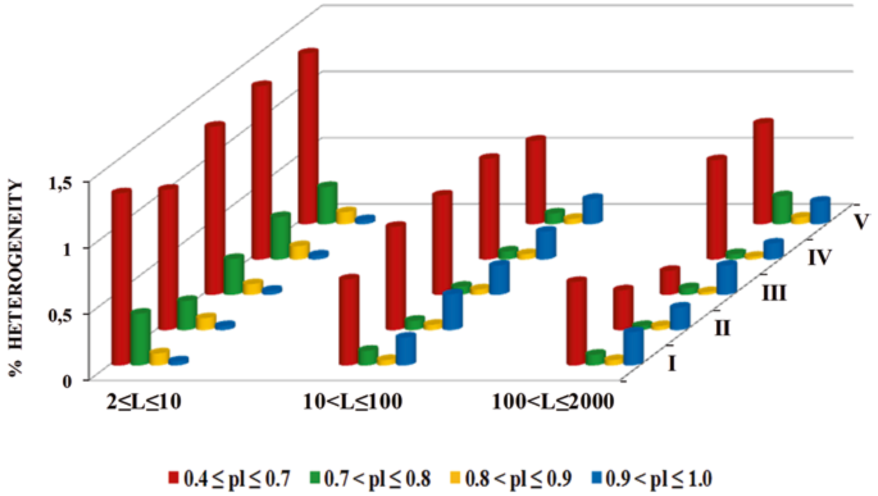

Fig. 3 Structural content for latent periodicity regions in genome of *A. thaliana* (the chromosomes I–V). Corresponding to revealed period *L*, for micro- $(2 \le L \le 10)$, mini- $(10 < L \le 100)$, and mega- $(100 < L \le 2000)$ satellites, coverage (as a percentage) of genome by periodicity (heterogeneity) regions with various preservation levels ($pl(L)$, *see* Eq. 1) is shown as separate *histograms*. The *columns in red* corresponds to highly divergent tandem repeats; that in *green* corresponds to moderately divergent tandem repeats; that in *yellow* corresponds to slightly divergent tandem repeats; and that in *blue* corresponds to perfect tandem repeats. See text for details

distinguished [18], as, for example, highly divergent micro-satellites in *S. cerevisiae* genome. The genomes of *A. thaliana* and *C. elegans* have similar composition of characteristic periodicities. Probably, sufficient percentage $(\sim 1.5\%)$ of mini- and mega-satellites is a consequence of active recombination processes [31–33] in the genomes of *Arabidopsis* and nematode. Domination of the micro-satellites in yeast genome could be related with the large number of genome replications in yeast growing and, consequently, with frequent replicase slippage [31–33] conducive to the elongation of such periodicity regions.

2.3.3 Revealing Latent Periodicity in the Genome Functional Regions

Using a link to the Sequence Viewer (http://www.ncbi.nlm.nih.gov/projects/sviewer/), for any periodicity region in the HeteroGenome database, one can receive information about the annotation of genome sequence, wherein the region is placed. As shown in the work [18], for the genomes of *S. cerevisiae*, *A. thaliana*, *C. elegans*, and *D. melanogaster*, correspondingly 80, 62, 65, and 67 % of the HeteroGenome groups (*see* Subheading 2.2) are placed in the genes. The rest of the groups from the database, practically, are situated in unassigned sequences of the genomes. However, it should be noted that 2.6 % of the groups from *D. melanogaster* genome is placed in the regions of various repeats.

2.3.4 Density of Distributing Latent Periodicity Regions Along the Chromosomes

How the latent periodicity regions are distributed over the chromosomes was studied for all genomes of model organisms in the database. Each chromosome was subdivided into sequential intervals of the same length, corresponding to 0.5% of chromosome total length. Then for each interval a summary length of the latent periodicity regions (total number of the nucleotides) revealed within the interval boarders was calculated. Such a value, normalized by total chromosome length and multiplied by 100%, was considered as a part (restricted by the interval) of the whole periodicity percentage on a chromosome. Summing the parts, over all intervals give an estimate of the whole periodicity percentage on a chromosome.

In investigating a density distribution within the intervals, only the group representatives from the HeteroGenome were considered, as corresponding to nonredundant estimate of chromosome coverage by the regions of latent periodicity. Besides, for every chromosome three additional distributions were obtained, corresponding to the density of micro- (period length is in a range $2 \le L \le 10$), mini- $(10 < L \le 100)$, and mega- $(100 < L \le 2000)$ satellites.

Investigation results for the density distribution of latent periodicity regions along the chromosomes are represented on the page Database Statistics (http://www.jcbi.ru/lp_baze/statistics/index.html) in the HeteroGenome. An example of such distributions for all chromosomes of *A. thaliana* is shown in Fig. 4.

Lengths of the unique sequential intervals for the chromosomes I-V were equal to 152138, 98491, 117299, 92925, and 134877 bp, correspondingly [17].

As one can see from the histograms in Fig. 4, the density distribution of the latent periodicity regions on chromosome is its unequivocal characteristic in genome. Such histograms can be considered as some kind of individual bar code for the chromosomes in genome.

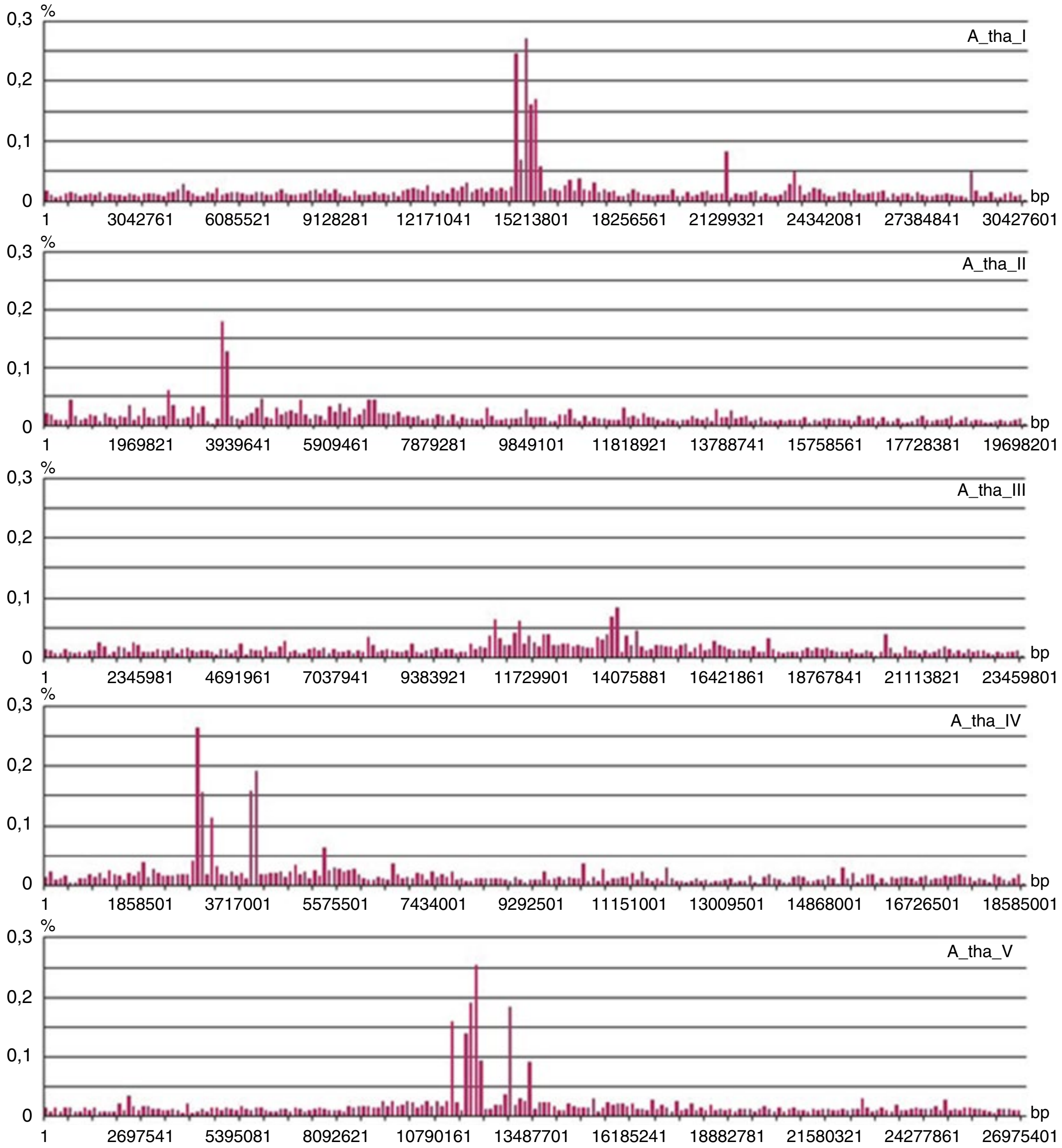

Fig. 4 Density distribution of latent periodicity regions along the chromosomes of *A. thaliana*. Height of an each column in histogram corresponds to percentage of local latent periodicity regions placed within a unique interval of chromosome division. See text for details

3 Spectral–Statistical Approach for Recognizing Latent Profile Periodicity

Initially, the 2S-approach was developed as complex of the methods searching for the regions of statistical heterogeneity in the genomes in order that further research of the regions will conduce to revealing new types of periodicity which are different from approximate tandem repeat. Among the HeteroGenome data, the sequences have been identified, wherein a new type of latent periodicity is recognized [18]. In the present section, new methods of the 2S-approach in recognizing such a type of latent periodicity, called latent profile periodicity or profility [20, 21], in DNA sequences are described.

3.1 Methodology of Recognizing Latent Profile Periodicity

Latent profile periodicity (latent profility) has a statistical basis. So, the statistical criteria which determine the similarity of analyzed DNA sequence with periodic random string of an etalon to recognize latent profility are formulated below. Consequently, a statistical hypothesis is tested that DNA sequence can be considered as a realization of etalon periodic random string. If such a hypothesis is accepted, existence of latent profile periodicity in DNA sequence is recognized, and a periodicity pattern is estimated. Hence, a special random string with periodicity pattern, consisting of independent random characters, is proposed as a model of the periodicity. This random string is perfect tandem repeat of such a pattern and called *a profile string*. The methods recognizing the latent profility are based on a model of profile string.

3.1.1 Model of Profile String and Notion of Latent Profile Periodicity

Profile string is a particular case of special random string which consists of independent random characters. In the general case, such a special random string of length n can be considered as a schema of the n independent tests of different random values, where each value has K outcomes as the letters of alphabet $A = \langle a_1, \ldots, a_K \rangle$. For DNA sequences $K = 4$ is the size of textual alphabet which is written as $A = \langle a_1, \ldots, a_4 \rangle = \langle a, t, g, c \rangle$. Every independent random value is called a random character, designated as $Chr(\mathbf{p})$ and determined by probability column $\mathbf{p} = \left(p^1, \ldots, p^K\right)^T$, where p^i is a probability of appearance for the ith $\left(i = \overline{1, K}\right)$ letter from the alphabet A. Consequently, such a schema of the n independent tests can be represented by formal string $Str_n(\mathbf{p}) = Chr(\mathbf{p}_1) \ldots Chr(\mathbf{p}_n)$. This string is n-dimensional random value, wherein $Chr(\mathbf{p}_j)$ is random character describing the jth $\left(j = \overline{1, n}\right)$ trial. Such a random string is unambiguously induced by a matrix $\Pi = \left(\mathbf{p}_1,, \ldots, \mathbf{p}_n\right) = \left(\pi^i_j\right)^K_n$ called n-profile matrix or profile matrix of the string $Str_n(\boldsymbol{\pi})$. In accordance to the works [12, 20–22], any integer number L out of a range $1 , \ldots, L_{\max}$, $L_{\max} \le \dfrac{n}{5K}$, is called *a test-period* for this string.

Let L be a test-period of the strings $Str = Str_n(\pi)\ 0 \le M < L$ and $Str_n(\pi) = Str_L(\pi_1)\ldots Str_L(\pi_m)Str_M(\pi_{m+1})$ is a decomposition of the string *Str* into the substrings of length *L*. If $M = 0$ ($\pi = (\pi_1,\ldots,\pi_m)$ and the string $Str_M(\pi_{m+1})$ is empty), then a matrix $\Pi_{Str}(L) = \dfrac{1}{m}\sum_{i=1}^{m}\pi$ is called *L-profile matrix of string* $Str = Str_n(\pi)$. If $M \neq 0$, then matrix $\Pi_{Str}(L)$ is corrected correspondingly. Thus, *a profile-matrix spectrum* Π_{Str}, determined at each test period, is introduced for the string $Str = Str_n(\pi)$. If $\pi_1 = \ldots = \pi_m = \pi_0$ and $\pi_0 = (\pi_{m+1}, \pi_{01})$, then string $Str_n(\boldsymbol{\pi})$ is called *L-profile string* with *a random periodicity pattern* $Ptn_L(\pi_0) = Str_L(\pi_0)$. Here, it is supposed that the pattern cannot be represented by consequent repeating of another random string. In this case a designation $Tdm_L(\boldsymbol{\pi}_0, n)$ is used for the string $Str_n(\boldsymbol{\pi})$. Besides, matrix $\boldsymbol{\pi}_0$ is called *a general profile matrix of string* $Tdm_L(\boldsymbol{\pi}_0, n)$, because this matrix induces a whole profile-matrix spectrum of the string. Integer L is called *a period length* of the string $Tdm_L(\boldsymbol{\pi}_0, n)$. If $L = 1$, then profile string $Tdm_1(\pi_0, n) = Tdm_1(\mathbf{p}, n) = \underbrace{Chr(\mathbf{p})\ldots Chr(\mathbf{p})}_{n\ \ times}$ will be called *a homogeneous string*, because its period length equals to unity.

Letter $a_i \in A$ can be identified with a random character which all components of probability (frequency) column are zeroes, excepting the *i*th unity component. Such a random character will be called *a textual character*. Consequently, any textual string in the alphabet A can be identified with corresponding special random string of the same length. Such a special string will be called *a textual string* also.

As for any random value for profile string $Str = Tdm_L(\pi_0, n)$, the n tests, corresponding to the string's scheme, can be carry out. In the result of these trials, a textual string *str* called a realization of the string $Str = Tdm_L(\pi_0, n)$ will be obtained. For the string *str*, one can pose a question on the existence of latent profile periodicity in it. If length n of the strings $Str = Tdm_L(\pi_0, n)$, $\left(L < L_{\max} \le \dfrac{n}{5K}\right)$, and *str* is sufficiently large, then their profile-matrix spectra will be statistically similar with great probability. This property is used in the 2S-approach for recognizing latent profile periodicity in the textual strings (DNA sequences).

In consistent with the 2S-approach, for recognizing latent profile periodicity in DNA sequence, it is necessarily to find such a profile string for that analyzed sequence can be considered as its realization. The search for such a profile string is carried out with the analysis of the spectral characteristics (the statistical spectra) of a textual string (DNA sequence) under consideration.

3.1.2 Methods for Estimating Period Length of Latent Profile Periodicity

To estimate the period of latent profile periodicity, the 2S-approach applies special statistical spectra of textual string which are introduced in the present section.

Let $Str = Str_n(\pi^*)$ be a random string of n independent random characters in the initial alphabet $A = \langle a_1,\ldots,a_K\rangle$. This string is induced by its n-profile matrix $\pi^* = (\mathbf{p}_1,\ldots,\mathbf{p}_n)$, where $\mathbf{p} = (p^1,\ldots,p^K)^T = \frac{1}{n}\sum_{j=1}^{n}\mathbf{p}_j = \Pi_{Str}(1)$ is a probability (frequency) vector of the letter (from the alphabet A) occurrence in the string $Str = Str_n(\pi^*)$. Then for each test-period λ of the string Str, λ-profile matrix $\Pi_{Str}(\lambda) = (\pi_j^i)_\lambda^K$ determines the following value $\Psi_1(\lambda)$:

$$\Psi_1(\lambda) = \Psi_1(\Pi_{Str}(\lambda),\Pi_{Str}(1),n) = \frac{n}{\lambda}\sum_{j=1}^{\lambda}\sum_{i=1}^{K}(\pi_j^i - p^i)^2 / p^i. \quad (5)$$

By such a way, for the string $Str = Str_n(\pi^*)$, a function $\mathbf{\Psi}_1$, defined at the test-periods of this string, is introduced that is called the string's *general spectrum*.

If $L \neq 1$, for nonhomogeneous profile string $Str = Tdm_L(\pi_0, n)$ (particularly, for textual tandem repeat), the following assertion can be mathematically strictly proven.

General spectrum $\mathbf{\Psi}_1$, *defined by Eq.* 5, *for nonhomogeneous profile string* $Str = Tdm_L(\pi_0, n)$ *has a period L. Maximal values of the spectrum* $\mathbf{\Psi}_1$ *are taken out only at the test-periods multiple of L. For homogeneous string* $(L = 1)$, *according to Eq.* 5, *its general spectrum takes on zero values.*

To visually illustrate the above assertions, Fig. 5 shows the graphics of general spectra for textual perfect tandem repeat (Fig. 5a) and profile string (Fig. 5b). This profile string is that its realizations are not the approximate tandem repeats.

By analogy with Eq. 5, for textual string *str* of length n, a general spectrum $\mathbf{\Psi}_1$ is introduced which at the test-period $\lambda < L_{\max} \leq \frac{n}{5K}$ takes on value:

$$\Psi_1(\lambda) = \Psi_1(\Pi_{str}(\lambda),\Pi_{str}(1),n) = \frac{n}{\lambda}\sum_{j=1}^{\lambda}\sum_{i=1}^{K}(\pi_j^i - p^i)^2 / p^i, \quad (6)$$

where $\Pi_{str}(\lambda) = (\pi_j^i)_\lambda^K$ is λ-profile matrix of the string *str* and $\Pi_{str}(1) = (p^1,\ldots,p^K)^T$. For the realizations of homogeneous string of length n, in accordance with Pearson goodness-of-fit test [29], a distribution of the $\Psi_1(\lambda)$ is statistically equivalent to the $\chi^2_{(K-1)(\lambda-1)}$ distribution, where χ^2_N is the χ^2-distribution with N degrees of freedom, i.e.,

$$\Psi_1(\lambda) \sim \chi^2_{(K-1)(\lambda-1)}. \quad (7)$$

In plotting a graph of general spectrum $\mathbf{\Psi}_1$ for textual string *str* obtained in the result of the realization of profile string

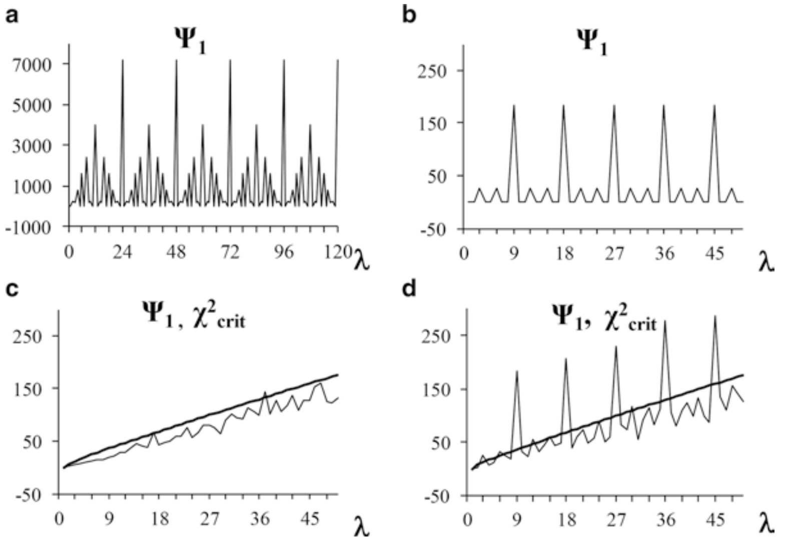

Fig. 5 General spectra (*thin lines*) of the profile and textual strings. (**a**) Perfect tandem repeat consisting of 100 copies of a pattern «atgcaattggccaaatttgggccc». (**b**) 9-profile string with general profile matrix, estimated over the string's "realization" (DNA sequence with general spectrum in (**d**). (**c**) Homogeneous (1-profile) string with the same base frequencies as in CDS (hsa:338872) from the KEGG database. (**d**) CDS of tumor necrosis factor-related protein (KEGG, hsa:338872, 1002 bp). *Bold line* in (**c**) and (**d**) shows a graphic of right-hand critical value $\chi^2_{crit}(N, \alpha)$. See text for details

$Str = Tdm_L(\pi_0, n)$, theoretical form of general spectrum $\mathbf{\Psi}_1$ for string $Str = Tdm_L(\pi_0, n)$ will be distorted. To illustrate such a distortion, the graphics of general spectra for a realization of homogeneous (1-profile) string (*see* Fig. 5c) and 9-profile CDS sequence (Fig. 5d) from the KEGG database [34] are shown. Furthermore, bold line in Fig. 5c, d shows a graphic of the right-hand critical value $\chi_{crit}{}^2(N, \alpha)$ correspondence to the test-period λ for the $\chi_N{}^2$-distribution at significance level $\alpha = 0.05$, where $N = (K-1)(\lambda-1)$.

According to Eq. 7, in the 2S-approach [20–22] for checking a hypothesis about homogeneity of textual string *str* (at significance level $\alpha = 0.05$), a spectrum $\mathbf{D}_1$ is used that at the test-period λ takes on value:

$$\mathrm{D}_1(\lambda) = \Psi_1(\lambda) / \chi^2_{crit}((K-1)(\lambda-1), \alpha). \tag{8}$$

If value $\mathrm{D}_1(\lambda) > 1$, then in accordance with goodness-of-fit test [29] at the test-period λ, heterogeneity is manifested in analyzed string. So, the $\mathbf{D}_1$ spectrum for textual string is called as *a spectrum of deviation from homogeneity.*

For nonhomogeneous profile string of length n, a probability distribution of the values in general spectra of the string's realizations at the test-period λ does not coincide with the χ^2-distribution, having $N = (K-1)(\lambda-1)$ degrees of freedom. In comparison with this χ^2-distribution, the existing distribution of the general spectrum values for the realizations of nonhomogeneous profile string induces essentially larger probability to exceed the critical level $\chi^2_{crit}((K-1)(\lambda-1), \alpha)$ than $\alpha = 0.05$. So, in the $\mathbf{D}_1$ spectra for the realizations of nonhomogeneous profile string, the test periods at which the values of the $\mathbf{D}_1$ spectrum exceed unity will be observed. In such a case textual string realizations will be called *heterogeneous strings.*

Figure 6a shows the $\mathbf{D}_1$ spectrum of deviation from homogeneity that was obtained from the $\mathbf{\Psi}_1$ general spectrum (*see* Figs. 5d or 6c). According to the $\mathbf{D}_1$ spectrum, human CDS (KEGG, hsa:338872) is considered as heterogeneous sequence.

The graphics of general spectra of profile string (Fig. 6d) and its "realization" (Fig. 6c) which in reality is CDS (KEGG, hsa:338872) from human genome are shown over again. As it follows from Fig. 6, the difference between the general spectra of profile string and its realization, practically, is of the form of graphic for a function linearly dependent on the test periods of the strings. Analogous to the $\chi^2_{(K-1)(\lambda-1)}$-distribution, with the increase of test-period λ, the freedom degrees of probability distribution for the values in the general spectrum $\mathbf{\Psi}_1$ of the original profile string

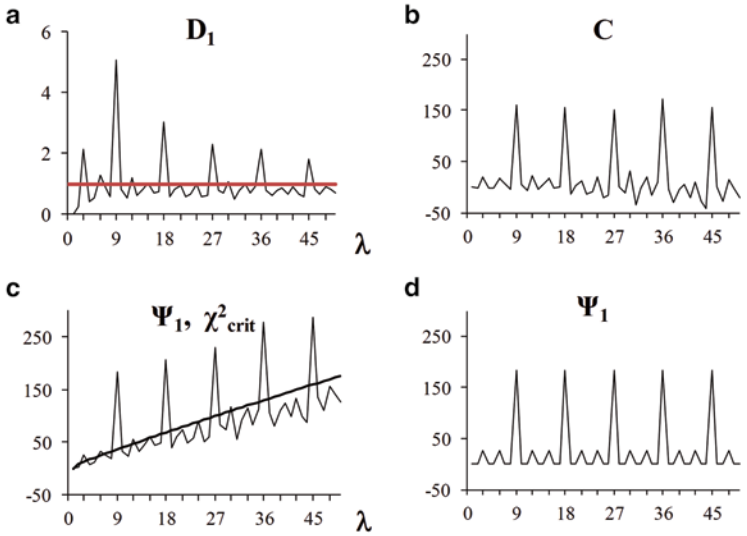

Fig. 6 The 2S-approach spectra for human CDS of tumor necrosis factor-related protein (KEGG, hsa:338872, 1002 bp). (**a**) Spectrum of deviation from homogeneity (*see* Eq. 8). (**b**) Characteristic spectrum (*see* Eq. 9). (**c**) General spectrum (*see* Eq. 6) is shown by thin line. *Bold line* draws a graphic of right-hand critical value $\chi^2_{crit}((K-1)(\lambda-1), 0.05)$. See text for details. (**d**) General spectrum of a profile string whose "realization" the analyzed CDS (KEGG, hsa:338872) can be considered

realizations ascend also. To level such a growth for realization *str*, a spectrum **C** is introduced as follows:

$$\mathrm{C}(\lambda) = \Psi_1(\lambda) - M\left(\chi^2_{(K-1)(\lambda-1)}\right) = \Psi_1(\lambda) - (K-1)(\lambda-1), \qquad (9)$$

where $M\left(\chi^2_N\right) = (K-1)(\lambda-1)$ is a mean value of the χ^2-distribution with N degrees of freedom. Further, the spectrum **C** is called *a characteristic spectrum of analyzed textual string*. The graphic of such a spectrum for an analyzed realization *str* is shown in Fig. 6b.

In comparing the characteristic spectrum (Fig. 6b) for the realization of an original 9-profile string with the general spectrum for 9-profile string (Fig. 6d), visual similarity both of the spectra is obvious. The 2S-approach is based on such a similarity in recognizing latent profile periodicity in the textual strings. For heterogeneous textual string realizations, a maximal value in characteristic spectrum is achieved (with allowance made to small random error) at a period of latent profile periodicity. Such the properties of characteristic spectrum are used in the 2S-approach for estimating period length of latent profile periodicity. For estimating period length in an analyzed textual string, the following rule is proposed.

At the beginning, a test-period L is selected out of string test periods at which the first clear-cut maximal value in characteristic spectrum **C** *is achieved. If $D_1(L) > 1$, then the test-period L is considered as an estimate of latent period of profile periodicity.*

Spectrum $\mathbf{D_1}$ of deviation from homogeneity is shown in Fig. 6a which has been obtained from the general spectrum $\mathbf{\Psi_1}$ (*see* Figs. 5d or 6c). Characteristic spectrum **C** (Fig. 6b) is corresponded to these spectra. According to the rule accepted above, an estimate of 9 bp is proposed as length of latent period of profile periodicity in analyzed coding DNA sequence (KEGG, hsa:338872) from human genome.

Efficiency of the rule formulated above for estimating period of latent profile periodicity in heterogeneous DNA sequences which cannot be considered as approximate tandem repeats has been proved in the works [20–22]. For such sequences, Fig. 7 shows the examples of characteristic spectra and spectra of deviations from homogeneity. It will be shown further that in these sequences the latent periodicities with the periods of $L = 10$ (Fig. 7a), $L = 84$ (Fig. 7c), and $L = 9$ (Fig. 7e) are revealed.

3.1.3 Pattern Estimate for Etalon of Latent Profile Periodicity on Basis of Goodness-of-Fit Test

For textual string *str*, an estimate of the period of latent profile periodicity $L > 1$ has been obtained basing on the **C** (*see* Eq. 9) and $\mathbf{D_1}$ (*see* Eq. 8) spectra of the string. Then by analogy with a general spectrum (*see* Eq. 6), to test whether the test-period L is a period of latent profile periodicity, the spectrum $\mathbf{\Psi_L}$ is used which at test-period λ takes on value:

$$\Psi_L(\lambda) = \Psi_L\left(\Pi_{str}(\lambda), \Pi_{Tdm_L}(\lambda), n\right) = \frac{n}{\lambda}\sum_{j=1}^{\lambda}\sum_{i=1}^{K}\left(\pi_j^{*i} - \pi_j^i\right)^2 / \pi_j^i \sim \chi^2_{(K-1)(\lambda-1)}, \qquad (10)$$

where $\Pi_{str}(\lambda) = \left(\pi_j^{*i}\right)_\lambda^K$ and $\Pi_{Tdm_L}(\lambda) = \left(\pi_j^i\right)_\lambda^K$ are λ-profile matrices of the textual string *str* and *L*-profile string $Tdm_L = Tdm_L\left(\Pi_{str}(L), n\right)$, correspondingly. For the realizations of *L*-profile string according to Pearson goodness-of-fit test [29], the following ratio is true:

$$\Psi_L(\lambda) \sim \chi^2_{(K-1)(\lambda-1)}. \qquad (11)$$

Using the statistics (Eq. 10) and the ratio (Eq. 11), the $\mathbf{D_L}$ spectrum of string *str* deviation from *L*-profility is introduced, taking (at the test-period λ) on the value:

$$D_L(\lambda) = \Psi_L(\lambda) / \chi^2_{crit}\left((K-1)(\lambda-1), \alpha\right), \qquad (12)$$

where $\chi^2_{crit}(N, \alpha)$ is a critical value of the $\chi_N{}^2$-distribution with N freedom degrees at significance level $\alpha = 0.05$. The $\mathbf{D_L}$ spectrum is used for checking a hypothesis about *L*-profility existence in analyzed textual string according to the following rule.

Let Q be a relative fraction of the test periods for an analyzed string at which the values of the $\mathbf{D_L}$ *spectrum are greater than unity. The hypothesis about L-profility existence in the string is accepted, if* $Q < 0.05$.

Let us give an example of how this rule is used. According to the spectra in Fig. 7, for three DNA sequences which are not approximate tandem repeats, the length estimates of 10, 84, and 9 bp have been proposed for the latent periods of profile periodicity. These estimates are visually confirmed in Fig. 8 with the help of the spectra of deviation from the corresponding profility.

The results of analysis for textual string *str*, where latent *L*-profile periodicity was revealed, allow supposing a random string $Ptn_L\left(\Pi_{str}(L)\right) = Str_L\left(\Pi_{str}(L)\right)$ of independent random characters as an estimate of this periodicity pattern. This random string is unambiguously characterized by profile matrix $\Pi_{str}(L)$ of string *str*. In this case a hypothesis about string *str* statistical similarity (at the significance level $\alpha = 0.05$) with profile string $Tdm_L(\Pi_{str}(L), n)$ is accepted. Thereby, profile string $Tdm_L(\Pi_{str}(L), n)$ is an etalon of profile periodicity for the string *str*. Besides, random string $Ptn_L(\Pi_{str}(L))$ is an estimate for pattern of this latent profile periodicity. Pattern $Ptn_L(\Pi_{str}(L))$ is an analogue of consensus-pattern deducing when approximate tandem repeats are recognized.

3.1.4 Methods, Reconstructing Spectrum of Deviation from Homogeneity and Confirming a Pattern Estimate for Etalon of Latent Profile Periodicity

Let a hypothesis about latent *L*-profility existence be accepted for heterogeneous textual string *str* (*see* Eq. 12 and text below). Consequently, the string *str* can be considered as a realization of *L*-profile etalon string $Tdm_L = Tdm_L\left(\Pi_{str}(L), n\right)$.

In forming etalon of profile periodicity $Tdm_L = Tdm_L\left(\Pi_{str}(L), n\right)$, goodness-of-fit test was used for an analyzed string *str*. But for obtained estimate of latent profile periodicity pattern, an additional

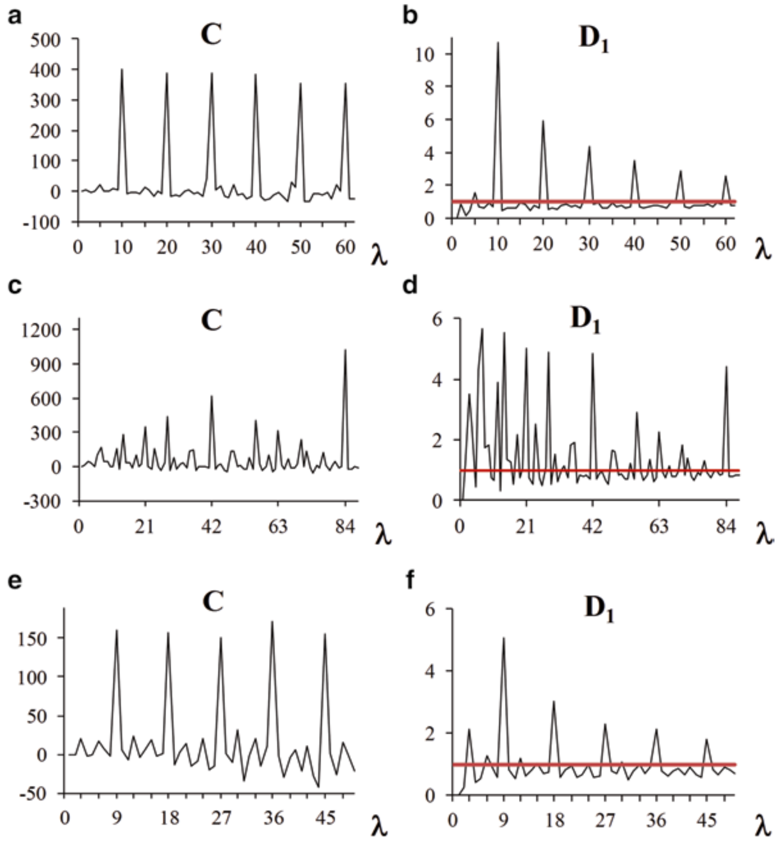

Fig. 7 The characteristic spectra **C** (**a**, **c**, **e**) and the $\mathbf{D_1}$ spectra of deviation from homogeneity (**b**, **d**, **f**) for the sequences that are not approximate tandem repeats. (**a**, **b**) Sequence on chromosome III of *C. elegans* (HeteroGenome, indices: 307381–308580, 1200 bp.). (**c**, **d**) CDS of human zinc finger protein (KEGG, hsa:26974, 1794 bp). (**e**, **f**) CDS of human tumor necrosis factor-related protein (KEGG, hsa:338872, 1002 bp)

conformation can be obtained. By analogy to the $\mathbf{D_1}$ spectrum (*see* Eq. 8), for random profile string Tdm_L, a spectrum $\mathbf{Th_L}$ is introduced, representing the string deviation from homogeneity, which at the test-period λ takes on value:

$$Th_L(\lambda) = \Psi_1\left(\Pi_{Tdm_L}(\lambda), \Pi_{Tdm_1}(\lambda), n\right) / \chi^2_{crit}\left((K-1)(\lambda-1), \alpha\right). \tag{13}$$

In fact, the $\mathbf{Th_L}$ spectrum is a theoretical reconstruction of the $\mathbf{D_1}$ spectrum for string *str*. To confirm an estimate of latent profile periodicity pattern, a method of comparing the spectra $\mathbf{D_1}$ and $\mathbf{Th_L}$ of deviation from homogeneity for the strings *str* and Tdm_L, correspondingly, was proposed in the works [20–22]. If for the string *str* a pattern estimate of latent profile periodicity etalon

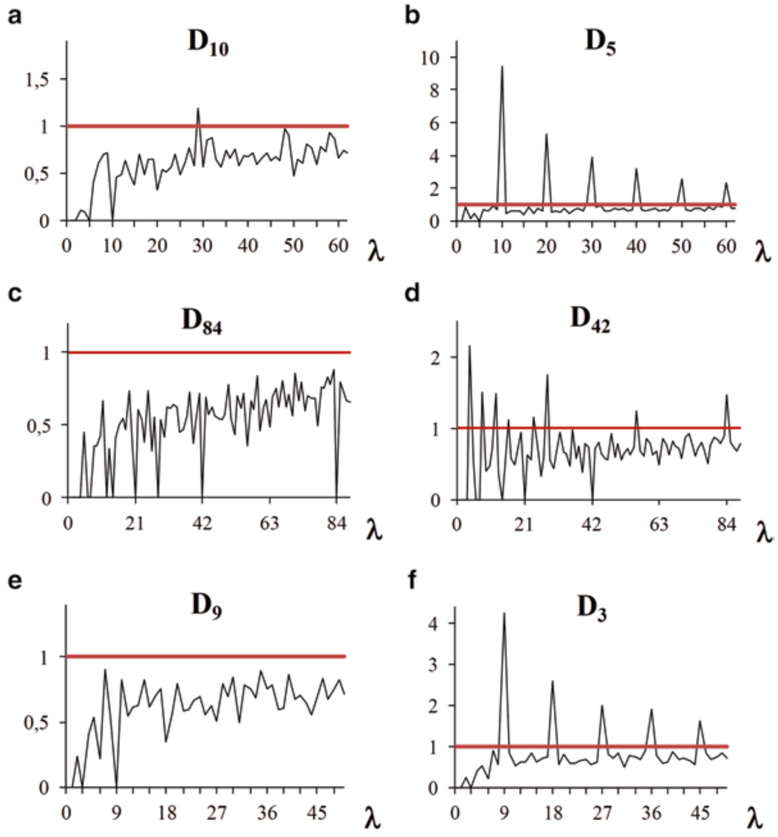

Fig. 8 Spectra of deviation from λ-profility ($\lambda = 10, 5, 84, 42, 9, 3$) for the following DNA sequences. (**a**, **b**) DNA fragment on chromosome III of *C. elegans* (HeteroGenome, indices: 307381–308580, 1200 bp); (**c**, **d**) CDS of human zinc finger protein (KEGG, hsa:26974, 1794 bp); (**e**, **f**) CDS of human tumor necrosis factor-related protein (KEGG, hsa:338872, 1002 bp)

$Tdm_L = Tdm_L\left(\Pi_{str}(L), n\right)$ is correct, then the spectrum $\mathbf{Th}_\mathbf{L}$ is obviously similar to the $\mathbf{D}_1$ spectrum. Figure 9d shows theoretical reconstruction of the $\mathbf{D}_1$ spectrum for human CDS (KEGG, hsa:26974). Visual similarity of this reconstruction with the original $\mathbf{D}_1$ spectrum of deviation from homogeneity (Fig. 9b) provides support for the revealed latent 84-profile periodicity.

3.2 Notion of 3-Regularity in Coding Regions of DNA Sequences

Earlier [21] in characteristic spectra of heterogeneous coding DNA sequences, regular repetition of the peaks at the test-periods multiple of three (*see*, e.g., Fig. 10a) was observed. Such a phenomenon contrary to the latent profility was called as 3-regularity of DNA sequences.

Let us describe a criterion of 3-regularity existence in DNA sequence [35]. Let us divide a range of definition for characteristic

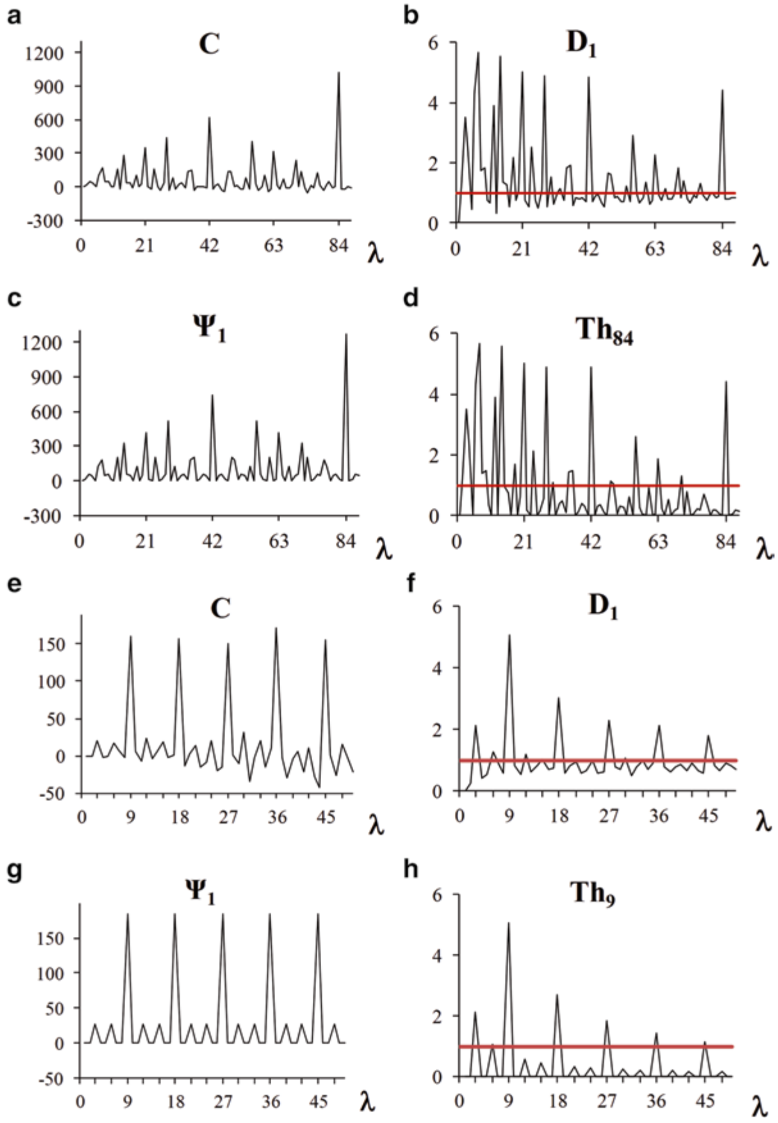

Fig. 9 Instantiation of pattern estimate for an etalon of latent profile periodicity with the help of the 2S-approach spectra for various DNA sequences. (**a**–**d**) CDS of human zinc finger protein (KEGG, hsa:26974, 1794 bp); (**e**–**h**) CDS of human tumor necrosis factor-related protein (KEGG, hsa:338872, 1002 bp)

spectrum of an analyzed DNA region into sequential triplets of the test periods. Within each triplet a test-period, corresponding to local maximal value in characteristic spectrum, is associated to unity, and the rest two test-periods are associated to zeros. As the result a binary string of the zeros and units is formed, i.e., textual string *str* in alphabet $A = \langle 0,1 \rangle$ of size $K = 2$. This string is compared with perfect periodic string of the same length and with periodicity pattern: 001. Index I_3, equal to a ratio of coinciding components between binary strings *str* and the perfect periodic one to the strings' length, is called an index of 3-regularity for analyzed sequence. If index $I_3 > 0.7$, then 3-regularity is observed in characteristic spectrum. For example, according to such a criterion in the characteristic spectra in Figs. 10a and 11b, d, f, corresponding to coding DNA sequences, 3-regularity is observed. In characteristic spectrum in Fig. 10b, corresponding to intron sequence,

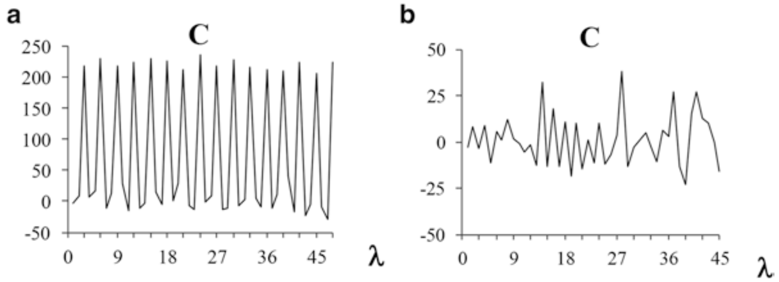

Fig. 10 Characteristic spectra of coding and noncoding DNA sequences: (**a**) Human transmembrane protein CDS (KEGG, hsa:80757, 960 bp); (**b**) Intron of human gene UCHL1 (ubiquitin carboxyl-terminal hydrolase isozyme L1) on chromosome IV (EID, INTRON_4 4383_NT_006238 protein_id:NP_004172.2, 917 bp.)

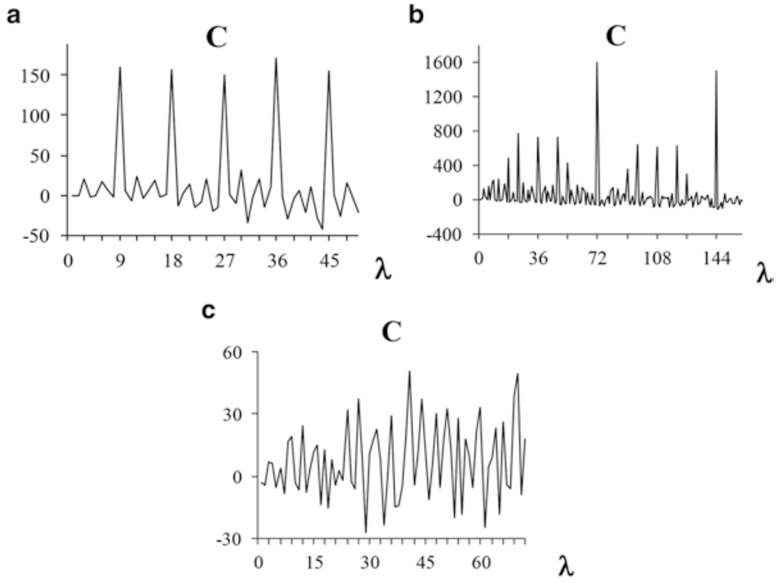

Fig. 11 Characteristic spectra of human CDSs from the KEGG database. (**a**) (hsa: 338872, 1002 bp) tumor necrosis factor-related protein; (**b**) (hsa:57055,1605 bp) deleted in azoospermia protein; (**c**) (hsa:149998, 1446 bp) lipase

3-regularity is not revealed, which is confirmed by the value of index $I_3 = 0.42 < 0.7$. In Figs. 10a and 11b, 3-regularity of the characteristic spectra is obvious. With the existence of 3-regularity in characteristic spectra in Fig. 11d, f is confirmed by the values of 3-regularity index $I_3 = 0.87$ and $I_3 = 0.78$, correspondingly.

3.3 Results of the 2S-Approach Application to Recognizing Latent Profile Periodicity and Regularity in DNA Sequences

Here, let us give a number of the examples of the 2S-approach application results for recognizing latent profile periodicity and 3-regularity in DNA sequences.

The methods of the 2S-approach revealed existence of latent profility of 33 bp (33-profility) in the genes of apolipoprotein family PF01442 from the Pfam (database of Protein families, http://pfam.sanger.ac.uk/) [36]. This family includes the apolipoproteins Apo A, Apo C, and Apo E which are the members of multigene family that, probably, has evolved from a common ancestor gene. Apolipoproteins perform lipid transport and serve as enzyme cofactors and the ligands of cellular receptors. The family amounts greater than 800 proteins from 100 different species. In Fig. 12a, b, c, the characteristic spectra of the coding regions of apolipoproteins for sea bream *Sparus aurata* (Apo A-I), chicken *Gallus gallus* (Apo A-IV), and mouse *Mus musculus* (Apo E) are shown. The maximal values in these spectra are achieved at test-periods

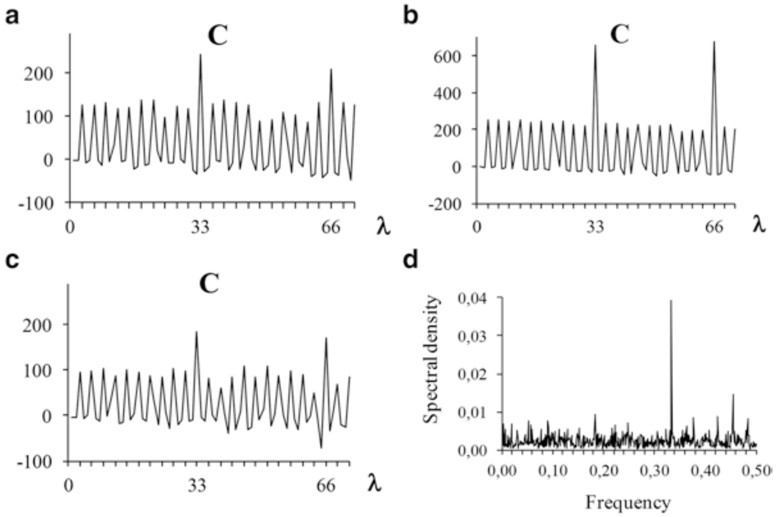

Fig. 12 Characteristic (**C**) and Fourier spectra for the coding regions in mRNAs of apolipoprotein family PF01442 (Pfam). (**a**) Apo A-I of *S. aurata* (GenBank AF013120, 34–816 bp); (**b**) Apo A-IV of *G. gallus* (GenBank Y16534, 37–1137 bp); (**c**) Apo E of *M. musculus* (GenBank M12414, 1–936 bp); (**d**) Fourier spectrum for the same sequence as in (**c**). Maximal peak in the spectrum is achieved at frequency 0.33, corresponding to regular heterogeneity of three bases

multiple of 33 bp. According to the 2S-approach, the latent 33-profility is recognized in these regions.

The well-known secondary structure of apolipoprotein family PF01442 consists of a few pairs of alpha-helix with 11 and 22 amino acid residues. Such a structure correlates with the profile periodicity of apolipoprotein genes of 33 bp. The peculiar pattern size of the latent profile periodicity in the genes of PF01442 family, possibly, influences on the formation of typical secondary structure in the protein family, and it is in agreement with the hypothesis about that family had originated from a common ancient gene.

In the characteristic spectra of coding regions, a regularity of the peaks at the test-periods multiple of three is observed (*see*, e.g., Fig. 12a, b, c). Thus, the first level of coding organization is manifested, that is, conditional by the genetic triplet code. Frequently, dominant peak in Fourier spectra at frequency 0.33 corresponds to this level (*see*, e.g., Fig. 12d). In existing 3-regularity, latent profility, which is distinct from 3-profility, reveals the second level in coding organization. Clear-cut maximal value in characteristic spectrum points at such level of the organization (Fig. 12a, b, c).

Existence of the latent 84-profility in coding DNA sequence (*see* Figs. 8c, and 9c, d) corresponds in protein to repeating zinc finger domain which includes one alpha-helix and two antiparallel beta-structures. As a rule, zinc finger domain counts about 20 amino acid residues, and it is stabilized by one or two zinc ions. DNA-binding transcription factors are the main group of the proteins with “zinc fingers.”

With the help of the 2S-approach, proposed methods search for 3-regularity and latent profility was done in 18140 human CDS from the KEGG database (Kyoto Encyclopedia of Genes and Genomes, http://www.genome.jp/kegg/) whose functional activity received experimental evidence. Within statistical errors of the methods, the CDSs are heterogeneous and 3-regular. Moreover, latent profile periodicity is observed for 74 % of the CDSs. The second level of encoding (different from 3-regularity and 3-profility) was revealed for 11 % of the analyzed CDSs, in that latent profility is displayed with period length multiple of three [21].

Analogous analysis was done for the introns also. The sequences of 277477 human introns (noncoding gene parts) from the EID (The Exon-Intron Database, http://utoledo.edu/med/depts/bioinfo/database) [37] were considered. Only 3 % of 3-regular sequences were revealed among them [21]. That is, in the frame of statistical method error, one can believe that the absence of 3-regularity is characteristic property for the introns.

4 Conclusion

Within the framework of the 2S-approach, the methods for recognizing two types of latent periodicity in DNA sequences were under consideration in the work. The first type was represented by the sequences which are similar to approximate tandem repeats. The second type is based on earlier introduced notion of latent profile periodicity (profility). The notion of latent profile periodicity generalizes notion of approximate tandem repeat. Presented methods of the 2S-approach allow recognizing these types in DNA sequences.

The application of the methods recognizing DNA sequences similar to approximate tandem repeats was demonstrated on the examples of genome analysis for model organisms from the HeteroGenome database. Special structure of the records in the HeteroGenome presents data on nonoverlapping latent periodicity regions on the chromosomes, providing with nonredundant data overview. The HeteroGenome database was design for molecular-genetic research and further study of latent periodicity phenomenon in DNA sequences. The analysis of data from the HeteroGenome has served to developing the spectral–statistical approach and passing on recognition of new type latent periodicity, called latent profile periodicity. Actuality of recognizing the latent profile periodicity due to such periodicity can correlate with the structural–functional organization of DNA sequences and their encoded proteins.

References

1. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27:573–580
2. Sokol D, Benson G, Tojeira J (2007) Tandem repeats over the edit distance. Bioinformatics 23:e30–e35
3. Issac B, Singh H, Kaur H, Raghava GPS (2002) Locating probable genes using Fourier transform approach. Bioinformatics 18:196–197
4. Sharma D, Issac B, Raghava GPS, Ramaswamy R (2004) Spectral Repeat Finder (SRF): identification of repetitive sequences using Fourier transformation. Bioinformatics 20:1405–1412
5. Paar V, Pavin N, Basar I, Rosandić M, Gluncić M, Paar N (2008) Hierarchical structure of cascade of primary and secondary periodicities in Fourier power spectrum of alphoid higher order repeats. BMC Bioinformatics 9:466
6. Wang L, Stein LD (2010) Localizing triplet periodicity in DNA and cDNA sequences. BMC Bioinformatics 11:550
7. Nunes MC, Wanner EF, Weber G (2011) Origin of multiple periodicities in the Fourier power spectra of the Plasmodium falciparum genome. BMC Genomics 12(Suppl 4):S4
8. Stoffer DS, Tyler DE, Wendt DA (2000) The spectral envelope and its applications. Stat Sci 15:224–253
9. Korotkov EV, Korotkova MA, Kudryashov NA (2003) Information decomposition method for analysis of symbolical sequences. Phys Lett A 312:198–210
10. Kumar L, Futschik M, Herzel H (2006) DNA motifs and sequence periodicities. In Silico Biol 6:71–78
11. Nair AS, Mahalakshmi T (2006) Are categorical periodograms and indicator sequences of genomes spectrally equivalent? In Silico Biol 6:215–222
12. Chaley M, Kutyrkin V (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

13. Salih F, Salih B, Trifonov EN (2008) Sequence structure of hidden 10.4-base repeat in the nucleosomes of C. elegans. J Biomol Struct Dyn 26:273–281
14. Epps J (2009) A hybrid technique for the periodicity characterization of genomic sequence data. EURASIP J Bioinform Syst Biol 2009:924601
15. Glunčić M, Paar V (2013) Direct mapping of symbolic DNA sequence into frequency domain in global repeat map algorithm. Nucleic Acids Res 41(1):e17
16. Gelfand Y, Rodriguez A, Benson G (2006) TRDB – The Tandem Repeats Database. Nucleic Acids Res 00(Database issue):D1–D8
17. Chaley MB, Kutyrkin VA, Tuylbasheva GE, Teplukhina EI, Nazipova NN (2013) Investigation of latent periodicity phenomenon in the genomes of eukaryotic organisms. Math Biol Bioinform 8:480–501
18. Chaley M, Kutyrkin V, Tulbasheva G, Teplukhina E, Nazipova N (2014) HeteroGenome: database of genome periodicity. Database article ID bau40
19. Epps J, Ying H, Huttley GA (2011) Statistical methods for detecting periodic fragments in DNA sequence data. Biol Direct 6:21
20. Chaley MB, Kutyrkin VA (2010) Structure of proteins and latent periodicity in their genes. Moscow Univ Biol Sci Bull 65:133–135
21. Chaley M, Kutyrkin V (2011) Profile-statistical periodicity of DNA coding regions. DNA Res 18:353–362
22. Kutyrkin VA, Chaley MB (2014) Spectral-statistical approach to latent profile periodicity recognition in DNA sequences. Math Biol Bioinform 9:33–62
23. Fields S, Johnston M (2005) Cell biology. Whither model organism research? Science 307:1885–1886
24. Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2015) GenBank. Nucleic Acids Res 43(Database issue):D30–D35
25. Boeva V, Regnier M, Papatsenko D, Makeev V (2006) Short fuzzy tandem repeats in genomic sequences, identification, and possible role in regulation of gene expression. Bioinformatics 22:676–684
26. Grover A, Aishwarya V, Sharma PC (2012) Searching microsatellites in DNA sequences: approaches used and tools developed. Physiol Mol Biol Plants 18:11–19
27. Gelfand Y, Hernandez Y, Loving J, Benson G (2014) VNTRseek – a computational tool to detect tandem repeat variants in high-throughput sequencing data. Nucleic Acids Res 42:8884–8894
28. Anisimova M, Pečerska J, Schaper E (2015) Statistical approaches to detecting and analyzing tandem repeats in genomic sequences. Front Bioeng Biotechnol 3:31
29. Cramer H (1999) Mathematical methods of statistics. Princeton University Press, Princeton, NJ
30. International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921
31. Dieringer D, Schlötterer C (2003) Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species. Genome Res 13: 2242–2251
32. Ellegren H (2004) Microsatellites: simple sequences with complex evolution. Nat Rev Genet 5:435–445
33. Richard GF, Kerrest A, Dujon B (2008) Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol Mol Biol Rev 72:686–727
34. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2012) KEGG for integration and interpretation of large-scale molecular datasets. Nucleic Acids Res 40(Database issue): D109–D114
35. Chaley M, Kutyrkin V (2016) Stochastic model of homogeneous coding and latent periodicity in DNA sequences. J Theor Biol 390:106–116
36. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR et al (2014) Pfam: the protein families database. Nucleic Acids Res 42(Database issue):D222–D230
37. Shepelev V, Fedorov A (2006) Advances in the Exon-Intron Database. Brief Bioinform 7: 178–185