Mathematical Biology and Bioinformatics
2018. V. 13. № S. P. t84-t103. doi: 10.17537/2018.13.t84

Translation of the original article
Chaley M.B., Kutyrkin V.A., Teplukhina E.I. et al. Mathematical Biology and Bioinformatics. 2013;8(2):480-501. doi: 10.17537/2013.8.480.

========== **BIOINFORMATICS** ==========

UDC: 577.322

Investigation of latent periodicity phenomenon in the genomes of eukaryotic organisms

Chaley M.B.[1*], Kutyrkin V.A.[2], Teplukhina E.I.[1], Tyulbasheva G.E.[1], Nazipova N.N.[1]

[1]*Institute of Mathematical Problems of Biology, Russian Academy of Sciences, Pushchino, Russia*
[2]*Moscow State Technical University n.a. N.E. Bauman, Moscow, Russia*

Abstract. Data analysis is presented for the HeteroGenome database first release which contains latent periodicity regions revealed in a number of eukaryotic organisms. Tandem repeats with different integrity of pattern copies, including the highly diverged repeats, have been identified in the genomes of *S. cerevisiae*, *A. thaliana*, *C. elegans* and *D. melanogaster*. Such data were obtained with the help of original spectral-statistical approach to searching for reliable regions of the latent periodicity in DNA sequences. Special structure of data presentation, consisting of the two levels, was proposed. On the first, nonredundant level the latent periodicity regions are considered as a whole and, additionally, on the second level only conservative elements of their periodic structures are shown. Such data presentation allowed estimating share of the periodicity regions as nearly 10% of the length in analyzed genomes. This estimate was deduced basing on the first level data. Quantitative and qualitative investigation of the latent periodicity regions, their divergence level over all chromosomes of the organisms considered, revealed characteristic types of periodicity in the genome of every organism. Histograms of density distribution for the latent periodicity regions on each chromosome of the genomes analyzed were obtained. Repertoire of period lengths were determinated. The HeteroGenome database has additional possibilities for inner data analysis and is accessible by URL: http://www.jcbi.ru/lp_baze/.

Key words: *latent periodicity, tandem repeats, genome analysis.*

INTRODUCTION

Tandem repeats (arrays of sequentially repeated copies of original DNA fragment or pattern) for a long time get the researchers' attention. On the one hand, such attention is caused by interest of understanding molecular mechanisms of rise and evolution of the repeats along with their functional significance in the genome, on the other hand – by possibility of new markers elaboration for the investigations in population and evolutionary genetics, basing on the repeats. Damage of pattern copies by point substitutions of original nucleotides as well as the insertions and deletions of both a single and a few nucleotides, leads to forming approximate tandem repeats. Point substitutions together with nucleotide insertions and

[*] maramaria@yandex.ru

deletions are commonly called mutations. Approximate tandem repeats with point nucleotide (nt) substitutions only are called widely as inexact tandem repeats. Approximate tandem repeats, including inexact ones, are the regions of latent periodicity in the genome.

Micro-satellites (with pattern of order of 10 nt) and mini-satellites (with pattern of order of 100 nt) are the most investigated group owing to their usage as genetic markers in medical forensics, for kinship assessment or positional cloning and in population and evolutionary genetics also [1].

It is considered that the micro-satellites arise in genome in the consequence of DNA polymerase "slippage" due to heat noise while working on periodic template [2]. In contrast to the micro-satellites, molecular mechanism leading to the longer mini-satellites arise and propagation is related with the processes of recombinant character such as unequal crossover or gene conversion [3]. Tandem repeats can also arise in the result of sequential gene duplication while homological recombining of the chromosomes occurs at meiosis stage [4].

Numerical tandem repeats are situated in centromeric and telomeric regions of the chromosomes [5]. Tandem repeats were found in fragile sites of the chromosomes [6,7]. There are instances when triplet microsatellites expansions in fragile sites caused human mental retardation [8]. Human neurological disorders may also be induced by "dynamic" mutations (contraction or expansion of the copies in tandem repeat) both in coding and noncoding regions, at that not only by triplet microsatellites mutations [9-11]. Tandem repeats in noncoding regions can affect to gene expression level, processes of transcription and translation [12].

A number of programs searching for perfect or nearly perfect tandem repeats: TRF [13], ACMES [14], MREPATT [15], STRING [16], mreps [17], ATRHunter [18] and others have been elaborated earlier. Different algorithms are the basis of these programs and, correspondingly, their results are not always coincident and depend on period length, copy number and divergence of repeat.

Over the last years the programs are elaborated which directed at finding of the more eroded tandem repeats and aimed at opportunity to study the repeats in evolutionary aspect also: TandemSWAN [19], IMEX [20], TRStalker [21], searching program based on a model of evolutive tandem repeats [22]. While perfect (or nearly perfect) tandem repeats vary in copy number and are dynamic owing to slippage mechanism in DNA replication, highly eroded approximate repeats are the more stable elements of the genome structure and their functional role is underexplored yet.

As a rule, the results of programs, searching for approximate tandem repeats, are used to develop various information resources as, for example, TRedD [23] database of human genome tandem repeats revealed with the help of algorithm [22]. The TRDB [24] is well-known database that includes the approximate repeats found by Repeats Finder (TRF) technique [13] in the sequenced genomes of prokaryotes and eukaryotes, particularly, in human genome. Database TRbase [25] interconnects the tandem repeats revealed in human genome by TRF technique [13] with localization of the genes on chromosomes, especially distinguishes the genes with defects causing genetic diseases.

It should be noted that proposed heuristic algorithms for finding out the highly eroded tandem repeats do not solve problem of reliability of the results obtained. To warrant revelation of approximate tandem repeat specifically, additional results' filtration is usually introduced. For example, in the program based on model of evolutive tandem repeats [22] pattern copy divergence level is restricted by the value of ~30%. Nevertheless, such value exceeds divergence level of 20%, chosen for probabilistic model in the TRF technique [13]. Besides, redundancy of the TRF technique has been noted earlier (there are the cases when a few patterns are proposed for the same DNA region) [19] along with instability of the technique results (In shifting at 1–3 nt, different pattern estimates are offered for the same DNA region.) [26].

In the present work original spectral-statistical approach [26–28] was used to search for reliable regions of the latent periodicity in the genomes of *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, *Caenorhabditis elegans* and *Drosophila melanogaster*. As it was shown earlier [26–28], such approach allows avoiding ambiguity, when latent periodicity structure is determined in approximate tandem repeats and allows optimizing an estimate of periodicity pattern. Theoretically, with the help of spectral-statistical approach [28] extremely eroded tandem repeats can be revealed for which degree of pattern copies' average divergence is of ~50%. General description of this approach is given in the next Section. The approach is used χ^2 statistics for testing homogeneity of DNA sequence at significance level what is characteristic for approximate tandem repeats which sequences are evidently heterogeneous. However, significant heterogeneity is necessary but not sufficient condition for determining periodicity of the types pointed above. Since the search results obtained in automatic regime are undoubtedly heterogenic sequences, but some additional analysis is needed for validation their periodical structure, further these results are mentioned as the heterogeneities in genome sequences. So, the name HeteroGenome (Heterogenic Genome) of database reflects this particularity of the data included.

The first release of the HeteroGenome database was not aimed at accumulation of data on the latent periodicity in all accessible now genomes of the organisms. First, we would like the data collected in the base will serve for insight of latent periodicity phenomenon in the genomes of living organisms, investigation of the phenomenon wide and, probably, its' purposive nature. So, in the present work quantitative and qualitative data analysis of the latent periodicity in the genomes of four organisms mentioned above was accentuated.

This article presents the first release of HeteroGenome database collecting the regions of latent periodicity revealed with the help of spectral-statistical approach [28] in various genomes. Owing to two-level data presentation, where the first nonredundant level includes nonintersecting sequences of the revealed latent periodicity regions, it becomes possible to estimate what is a percent of these regions from length of the genome under consideration. Moreover, with the help of special parameter, reflecting average safety level for pattern copies in repeat, which is described in the next Section, one can analyze quality of periodic structure in each region. Such analysis allowed revealing characteristic types in every one of the genomes considered in the present work. By this way the HeteroGenome database may be useful both for the investigations in functional and evolutionary genomic areas, in searching for tandem repeats which are characteristic for analyzed genome, and for latent periodicity phenomenon study in DNA sequences. Database has user friendly interface, various options for additional data analysis and allows uploading query result also. The HeteroGenome database is freely available at URL http://www.jcbi.ru/lp_baze/.

MATERIALS AND METHODS

In the first release of HeteroGenome database the latent periodicity regions were searched for in four genomes of well-studied model organisms [29]: *S. cerevisiae*, *A. thaliana*, *C. elegans* и *D. melanogaster*. The genomes were sequenced among the first ones and are sufficiently exact and annotated DNA sequences. As well, they present eukaryotic genomes ranging from single cellular (baker yeast) to multicellular organisms of plants (arabidopsis) and animals (nematode) that conduce general consideration of latent periodicity phenomenon in the genome.

DNA sequences of the whole genome length were obtained from the site ftp://ftp.ncbi.nih.gov/genomes/. Data on the chromosomes of organisms analyzed are shown in Table 1.

Table 1. Source references for the genomes of model organisms

Organism	Chromosomes	GenBank Identifiers	
S. cerevisiae	I	NC_001133.7	GI:144228165
	II	NC_001134.7	GI:50593115
	III	NC_001135.4	GI:85666111
	IV	NC_001136.8	GI:93117368
	V	NC_001137.2	GI:7276232
	VI	NC_001138.4	GI:42742172
	VII	NC_001139.8	GI:162949218
	VIII	NC_001140.5	GI:82795252
	IX	NC_001141.1	GI:6322016
	X	NC_001142.7	GI:116006492
	XI	NC_001143.7	GI:83722562
	XII	NC_001144.4	GI:85666119
	XIII	NC_001145.2	GI:44829554
	XIV	NC_001146.6	GI:117937805
	XV	NC_001147.5	GI:84626310
	XVI	NC_001148.3	GI:50593503
	MT	NC_001224.1	GI:6226515
A. thaliana	I	NC_003070.9	GI:240254421
	II	NC_003071.7	GI:240254678
	III	NC_003074.8	GI:240255695
	IV	NC_003075.7	GI:240256243
	V	NC_003076.8	GI:240256493
C. elegans	I	NC_003279.4	GI:86561680
	II	NC_003280.4	GI:86562519
	III	NC_003281.5	GI:86563600
	IV	NC_003282.3	GI:72185816
	V	NC_003283.5	GI:86564547
	X	NC_003284.5	GI:86565306
D. melanogaster	2L	NT_033779.4	GI:116010444
	2R	NT_033778.3	GI:116010442
	3L	NT_037436.3	GI:116010443
	3R	NT_033777.2	GI:56411841
	4	NC_004353.3	GI:116010290
	X	NC_004354.3	GI:116010291

In the present work original spectral-statistical approach [26–28] is employed to search for approximate tandem repeats. In essence, latent periodicity revelation with the help of the approach is oriented at significant heterogeneity detection at the test-periods of analyzed nucleotide sequence. For each test-period L analyzed sequence is divided into the substrings of length L (last substring may have the smaller length). If n is length of analyzed sequence than $R_L = n/L$ is test-exponent for the test-period L. Such division into the substrings allows calculating frequency π_j^i of the ith character from nucleotide sequence alphabet in the jth position of test-period. Matrix $\pi = (\pi_j^i)_L^K$ is called a sampled L-profile matrix for analyzed sequence, where K is size of nucleotide sequence alphabet. Frequency p^i of the ith alphabet character in analyzed sequence is determined according to matrix $\pi = (\pi_j^i)_L^K$ as following:

$$p^i = \frac{1}{L}\sum_{j=1}^{L}\pi_j^i, \quad i = 1,...,K. \qquad (1)$$

To check homogeneity of sequence at test-period L normalized Pearson χ^2-statistics [28] is used.

$$\nu_{NP}(L,n) = R_L \sum_{j=1}^{L}\sum_{i=1}^{K} (\pi_j^i - p^i)^2 \big/ p^i(1-p^i)\,. \qquad (2)$$

As tandem repeats are searched for in the nucleotide databases of large volume significance level (type I error) $\alpha = 10^{-6}$ was chosen to check homogeneity of DNA sequences. Critical value $\chi^2_{crit}(\alpha, N)$ with $N = (K-1)(L-1)$ freedom degrees is correspond to fixed value L of test-period. Therefore, if a value of statistics $\nu_{NP}(L,n)$ for analyzed string of length n at test-period L meets the condition

$$\nu_{NP}(L,n) \big/ \chi^2_{crit}(\alpha,(K-1)(L-1)) \le 1\,, \qquad (3)$$

than hypothesis of the string homogeneity at test period L is accepted, otherwise the string is considered heterogenic. So, the following function H is used as spectral characteristics of analyzed nucleotide string

$$H(L) = \nu_{NP}(L,n) \big/ \chi^2_{crit}(\alpha,(K-1)(L-1))\,, \qquad (4)$$

where $L = 1,...,L_{\max}$ $(L_{\max} \sim n/5K)$.

Graphic of function H (H-spectrum) which is called *spectrum of heterogeneity manifestation* in analyzed sequence clearly demonstrates manifestation of significant string heterogeneities at those test-periods, where $H(L) > 1$. Such the test-periods form sequence's *spectrum of heterogeneity structure*, that analyzed further with the help of additional spectral-statistical parameter.

At each test-period L of analyzed sequence according to the sampled L-profile matrix $\pi = (\pi_j^i)_L^K$ a value is calculated for parameter

$$pl(L) = \frac{1}{L}\sum_{j=1}^{L} \max\{\pi_j^i : i \in 1,...,K\}\,, \qquad (5)$$

what is called a *character preservation level* at test-period L.

Hence, the spectrum of heterogeneity structure is analyzed with the help of spectrum of character preservation level (pl-spectrum) in the sequence considered. That test-period in the spectrum of heterogeneity structure, which is pointed at by the first maximum in the spectrum of character preservation level, is considered as size estimate of periodicity pattern in approximate tandem repeat (see fig. 1). Such maximal value of the pl-spectrum may be interpreted as average preservation index of periodicity pattern copies in the repeat. Figure 1 is demonstrative example of how combined usage both of two parameters (H-spectrum and pl-spectrum) allows unambiguously estimating periodicity pattern size. In the sequence analyzed in fig. 1 H-spectrum identifies heterogeneity structure at the test-periods multiplied by seven. Maximum of the pl-spectrum distinguishes among them test-period of 21 nt accepted as estimate of periodicity pattern size.

Problem of the results reliability for revealing approximate tandem repeats in the conditions of small statistical samples (when number of pattern copies in tandem repeat is sufficiently small) in a frame of spectral-statistical approach is solved basing on stochastic model of heterogeneity manifestation in textual strings [28]. This model allows using additional statistical tests while checking the hypothesis about heterogeneity existence in DNA sequence.

As from algorithmic point of view, latent periodicity revelation is considered very difficult problem if period is not a priory known, so to find out the periodicity in DNA sequences,

technique has been chosen that reveals the regions of highly significant (at level the $\alpha = 10^{-6}$) heterogeneity with the help of series of overlapping windows, where each window scans analyzed sequence by shifting at changeable step. The smallest window length is equal to 30 nt, and length of each following window becomes twice more. So, general strategy of program complex, realizing spectral-statistical approach [28] for searching the approximate tandem repeats, is likely shotgun-strategy of the genome sequencing [30]. In frame of such strategy, firstly short and overlapping fragments are sequenced, then their computer agglomeration is done that forms the longer regions. Analogous to this procedure, originally obtained data about the regions of significant heterogeneity in the genomes of model organisms considered, were additionally processed and the boarders of revealed regions of heterogeneity were optimized.

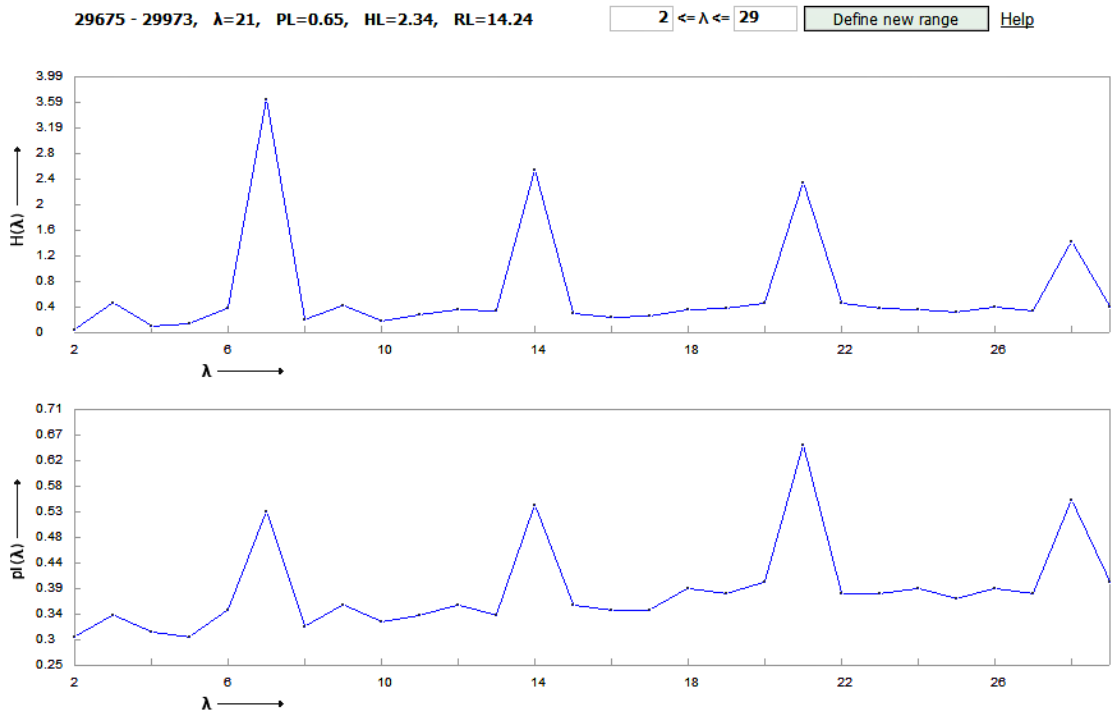

Figure 1. Spectral characteristics of DNA sequence on chromosome V (29675–29973 nt) from *C. elegans* genome in the HeteroGenome database. At the top: Spectrum of heterogeneity manifestation (H -spectrum, see Eq. (4)). At the bottom: Spectrum of character preservation level (pl -spectrum, see Eq. (5)). Maximal peak of the pl -spectrum corresponds to the latent profile periodicity pattern length of 21 nt.

RESULTS

Application of the spectral-statistical approach [26–28] allowed to create program complex, that reliably reveals the regions of latent periodicity including eroded and fuzzy tandem repeats. Theoretical limitary level of divergence for pattern copies in tandem repeats which are revealed by the complex is equal to 50%. Complex of the programs developed reveals both micro-satellites with minimal pattern length of 2 nt and mini-satellites with pattern of ~ 10 – 100 nt along with mega-satellites having pattern length ranging from 100 nt to 2000 nt. Minimal quantity of pattern copies revealed in approximate tandem repeats equals two. Search results for the regions of latent periodicity in the complete genomes of model organisms *S. cerevisiae*, *A. thaliana*, *C. elegans* and *D. melanogaster* after passing through special procedures, optimizing the region boarders, have been collected in the HeteroGenome database (http://www.jcbi.ru/lp_baze). The results of analysis for data collected in the database are shown on page “Database Statistics” (see fig. 2).

The HeteroGenome is relational database governed by DBMS MySQL. Searching system over all fields (see fig. 2) is organized with possibility of data sorting for any field (Region Location, Region Length, Period Length, Exponent, Preservation Level). Detailed description of all fields and their possible values is outputted in different windows opened in response of cursor setting on field name. User manual with the examples demonstrating how to operate with the HeteroGenome is accessible in the database site. There is possibility to upload search results as textual file also. In accordance with two-level logical record structure (see fig. 3), that will be described below, the HeteroGenome interface offers to choose informational request level: the first nonredundant and the second general (simple) level, where a list of all sequences relevant to request are shown.

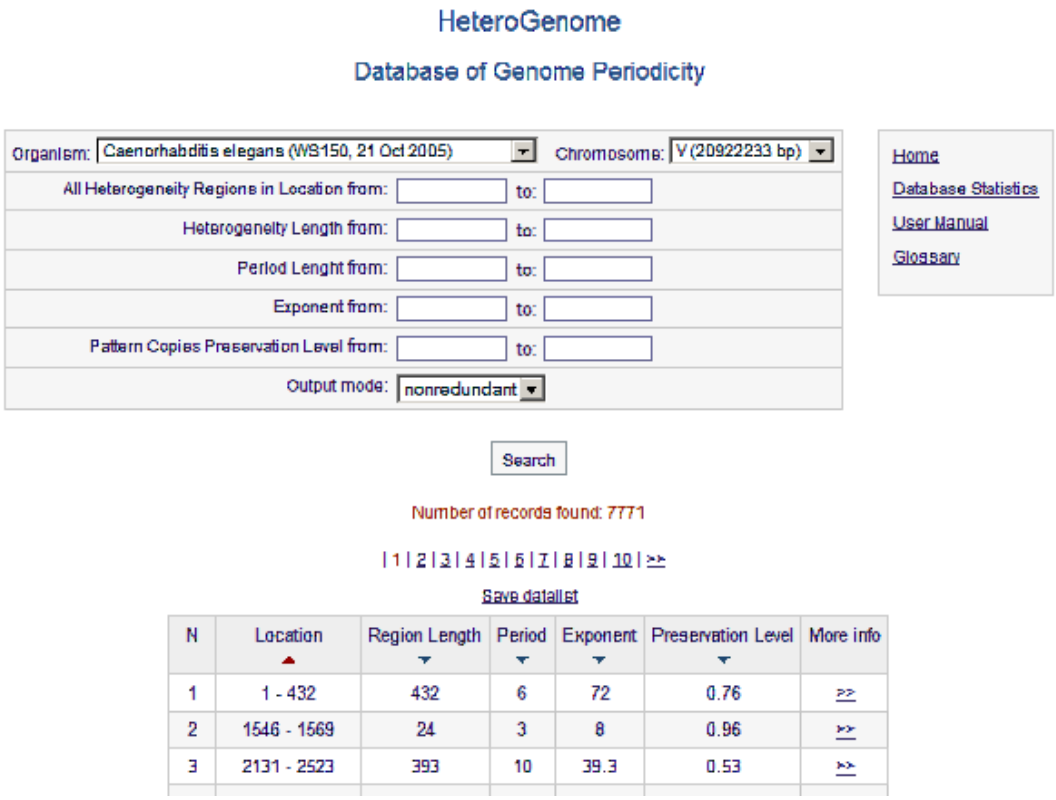

Figure 2. User request and data output in the HeteroGenome are shown. All latent periodicity regions found out on chromosome V from *C. elegans* genome, corresponding to the first nonredundant level of the records, are displayed.

For each sequence from the HeteroGenome database one can view of *H*- and *pl*- spectra (see Equations (4), (5) and fig. 1) in different windows (see fig. 3) and basing on the spectra periodicity pattern size estimate is proposed. Furthermore, there is possibility of viewing the sequence as profile, i.e. presented by column of the sequential segments of length equal to pattern size estimate (see fig. 4).

Being guided by information obtained from the spectra analysis, user may vary sequence segment length to define pattern size in approximate tandem repeat more exactly. Applying built-in graphical interface Sequence Viewer (http://www.ncbi.nlm.nih.gov/projects/sviewer/) (see fig.3), user may get information about region considered localization and it's functional context on chromosome.

For data nonredundant presentation in the HeteroGenome database one logical record corresponds to a group of DNA sequences from the same chromosome, if the sequences are intersecting and have statistically significant heterogeneity (latent periodicity) and (or) have the same or multiplied period length. Two level of data presentation are determined for such a group. The longest DNA sequence in group is outlined as the group representer, that specifies

the first level. The rest sequences are attributed to the second level. As a rule, they correspond to well determinated local periodicity structures in the sequence-representer. Example of such two-level record organization in the HeteroGenome is shown in figure 3. The values of parameters for group representer (sequence of the first level) are shown in the upper table. Below the title INTRINSIC HETEROGENEITIES, in another table, the parameter values for the sequences of the second level are presented. Visual presentation of the whole group is given on the scaled scheme at the bottom.

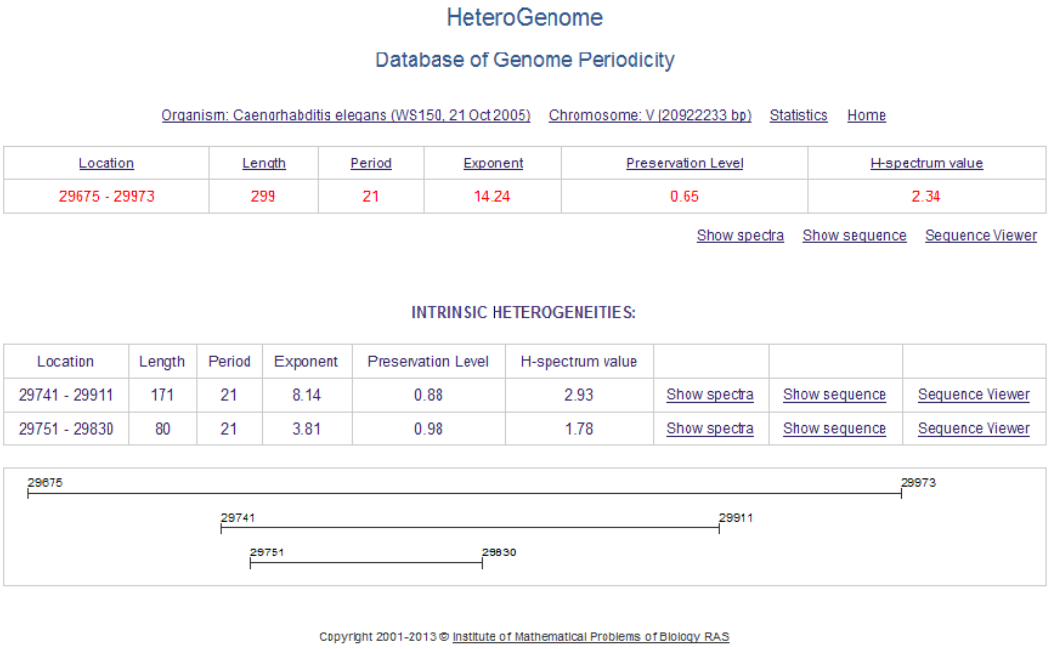

Figure 3. Two-level structure of record in the HeteroGenome database is presented. DNA sequences of *C. elegans* genome located on chromosome V within the region 29675 – 29973 nt, where latent periodicity of 21 nt is revealed, constitute a group divided in two levels. The first level is associated with the longest sequence (for which parameters are shown in red) called group representer. The second level of the group (INTRINSIC HETEROGENEITIES) is formed by the sequences of inner periodic structures found in the group representer sequence. Graphical scheme of the whole group is shown at the bottom. The first level sequences from all groups for each genome in the HeteroGenome database constitute nonredundant data on latent periodicity in the genome.

There are many groups in the database including the sequence representer only. Practically, the groups do not intersect each other. So, the sequences of representers form nonredundant chromosome coverage by the regions of significant heterogeneities (latent periodicities).

Search of information about the latent periodicity in the HeteroGenome (with given period length, size of periodicity region, preservation level of pattern copies and others), as will be shown further, may be done both through the first nonredundant level and through all the sequences at both levels.

Period length and the coordinates of latent periodicity (heterogeneity) region in the HeteroGenome database may be determined more exactly in result of visual analysis of spectral parameters (see fig. 1), DNA sequence division into the segments of proposed period length (see fig. 4) and while giving a length of flanking regions of original sequence also.

In some cases, additional view of group sequences facilitates more correct reading of data whose analysis and distribution over the groups have been done by computer programs. After group content has been analyzed, user may revise the group size or divide it into independent subgroups.

Analysis of latent periodicity in the HeteroGenome

Comparison of data for the genomes of *S. cerevisiae*, *A. thaliana*, *C. elegans* and *D. melanogaster* from HeteroGenome database with corresponding data from TRDB database [24] has shown that HeteroGenome collects practically all the same tandem repeats presented in the TRDB and, moreover, supplements the repeats with data on highly eroded tandem repeats.

In investigating evolution and functional significance of the latent periodicity regions in the genome, a share of the whole genome length covered by such the regions is important quantitative index. Nonredundant data on reliable heterogeneity (latent periodicity) in the HeteroGenome database allow estimating percent of the tandem repeats (including highly eroded tandem repeats) in the genomes of model organisms precisely enough. Table 2 shows such the estimates.

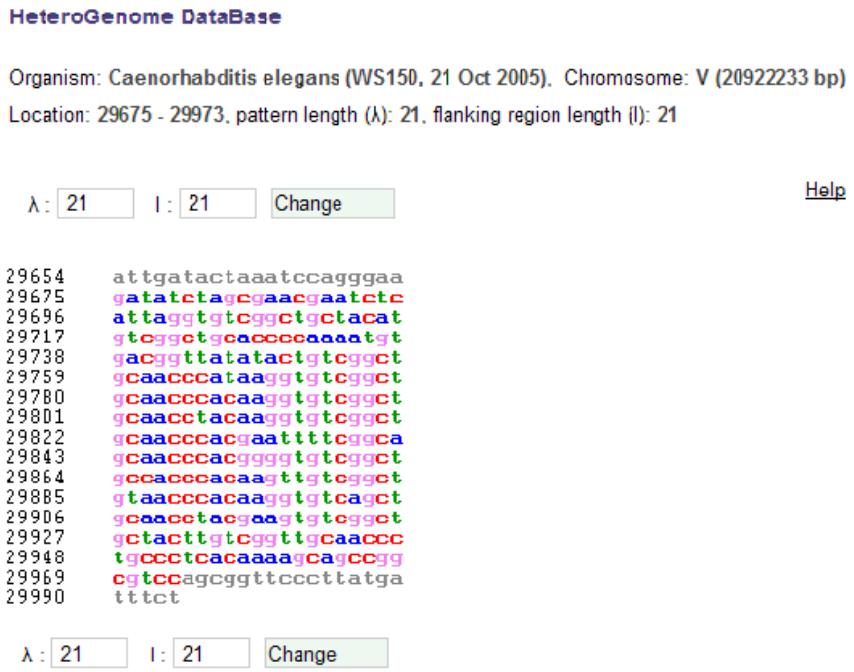

Figure 4. Example of profile for analyzed sequence in the HeteroGenome database. Division into the segments for DNA sequence from chromosome V of *C. elegans* genome (29675 – 29973 nt) is done in according to the length of latent periodicity pattern λ = 21 nt (see fig. 1). Characters of the latent periodicity region are presented by colors while the ones of flanking region are shown in gray. Length of flanking regions is set equal to pattern size ($l = \lambda$) by default.

Table 2. Share of heterogeneity (latent periodicity) regions in the genomes of analyzed model organisms.

	Genome length, nt	Total length of revealed heterogeneity regions, nt	Length of heterogeneity regions / Genome length, %
S. cerevisiae	12070900	419909	3.5
A. thaliana	119146348	4247672	3.6
C. elegans	100269917	6692629	6.7
D. melanogaster	120381546	5108483	4.2

As it will be shown further, micro- and mini-satellites (period length is less than 100 nt) constitute the most part of periodicity regions in the genomes of organisms analyzed. It is known that in human genome such the regions constitute 3% [30]. All together with other tandem repeats (which period length is in order of 1000 nt) they amount about 10% [19] of genome length. Considering data of Table 2 also, it could be supposed that periodicity in eukaryotic genome varies within 10%. Possibly, such percent is due to a balance of molecular mechanisms of tandem repeats origin and divergence of their sequences that stabilizes length of the repeats.

Influence of latent periodicity at chromosome length

Periodicity regions are unstable areas in the genome that can both expand and diminish in length due to the mechanisms of slippage of DNA replicase, recombination and duplication [2-4]. Mutations (point nucleotide substitution, insertions/deletions of a few nucleotide) disturb in time determinated structure of DNA periodicity regions, at that stabilizing their length. As the method used in the work for revealing approximate tandem repeats allows nonredundant estimation of the repeats' share in the genome, so influence of periodicity regions' evolution at the chromosomes may be investigated. Let us consider a share of the periodicity regions (reliable heterogeneity presented by tandem repeats) in relation to chromosomes' length of analyzed model organisms (see fig. 5).

For each model organism characteristic scattering of the values for percent of coverage by periodicity regions on the chromosomes is considered. The most difference (4.95%) between the maximal (8.91% for chromosome I) and minimal (3.96% for chromosome X) values is observed in *C. elegance* genome. Scattering of the percent values in the genome of *D. melanogaster* (3.11%) is approximately by one third less, ranging from 3.30% (chromosome IV) to 6.41% (chromosome X). For both organisms such interval of the values is due to particularities of X chromosome in the genomes, as for *C. elegance* this chromosome has minimal percent of coverage but for *D. melanogaster* X chromosome has maximal percent of coverage among the rest chromosomes in the genome. Interval of the percent values for coverage of the chromosomes by periodicity regions (reliable heterogeneities regions) in *D. melanogaster* genome is comparable with analogous interval in *S. cerevisiae* genome. Last interval equals to 3.57% between the maximal (6.27% for chromosome I) and minimal (2.7% for chromosome XVI) values. Let us note, that chromosome I, leading by periodicity percent in *S. cerevisiae* genome, is the shortest chromosome in yeast genome.

If in the genomes of *S. cerevisiae*, *C. elegans* and *D. melanogaster* interval of the values of periodicity percent over the chromosomes is comparable with the mean value for individual genome, then in the genome of *A. thaliana* such the interval is no more than 0.75%. As one can see in fig. 5 for arabidopsis, while length of the chromosomes is growing, percent of periodicity regions (heterogeneity regions) is keeping practically the same. Generally, with growth of chromosome length, percent values of periodicity regions are prone to stability or diminishing for all analyzed genomes of model organisms. One may think that due to instability of tandem repeats and their capability to elongation in consequence of the mistakes during replication they have influenced at growth of chromosome length moderately (~10%) but rather noticeably.

Analysis of periodic structure conservation in heterogeneity regions

According to the HeteroGenome database, series of the figures 6 – 9 presents the histograms of the revealed latent periodicity regions distribution with correspond to preservation level (*pl*-parameter, see Equation (5)) of their periodic structure. For each group of micro- (period length $2 \le L \le 10$), mini- (period length $10 < L \le 100$) and mega- (period length $100 < L \le 2000$) satellites percentage of chromosome occupied by highly eroded ($0.4 \le pl \le 0.7$), eroded ($0.7 < pl \le 0.8$), slightly eroded ($0.8 < pl \le 0.9$) and perfect

($0.9 < pl \leq 1.0$) tandem repeats of such kind is shown. Chromosomes in the figures 6 – 9 are arranged in order of the histograms' similarity.

As one can see in fig. 6, regions of the latent periodicity in *S. cerevisiae* genome are mainly presented by highly eroded sequences of micro-satellites with percentage ~2%. Highly eroded mini-satellites constitute less than 1% in this genome. While excluding chromosome I

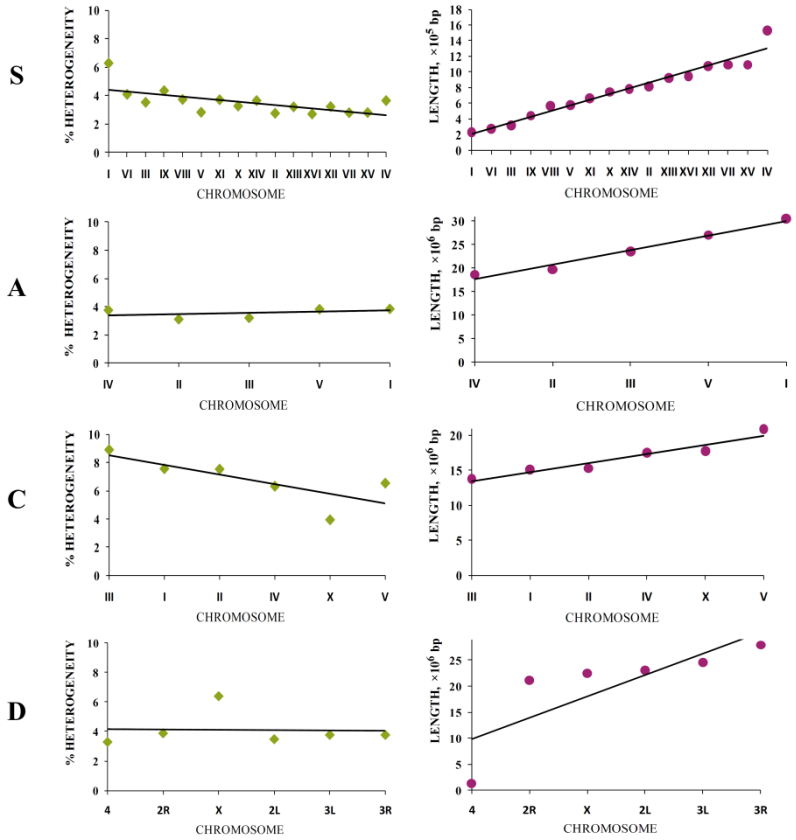

Figure 5. Percent of coverage by the regions of latent periodicity (heterogeneity) for the chromosomes of model organisms (**S**) *S. cerevisiae*, (**A**) *A. thaliana*, (**C**) *C. elegans* and (**D**) *D. melanogaster*. Chromosomes for each organism are arranged by growing length, as show in the graphics on the right. Trends are shown by solid lines. Percentage of the latent periodicity regions on the chromosomes was estimated according to nonredundant level of the records in HeteroGenome database.

and chromosome II, mega-satellite repeats in *S. cerevisiae* genome with period length of more than 100 nt are practically absent. Some chromosomes in *S. cerevisiae* genome have similar histograms in all three satellite groups, for example, chromosomes VIII and XII, chromosomes XIII, VI, XIV and chromosomes XI, XV, VII also.

For plant *A. thaliana* (fig.7) and round worm *C. elegans* (fig. 8), whose genome lengths are in order of magnitude more than length of yeast *S. cerevisiae* genome, the histograms presented disclose similar tendencies of qualitative contents of tandem repeats in genome. First, highly eroded mini-satellites constitute in both the genomes significant part ~1–1.5% comparable with micro-satellites' percentage. Hence, micro- and mini-satellites of *A. thaliana* and *C. elegans* input similar in structural and functional genome organization. Second, share of mega-satellite repeats which is practically absent in yeast genome (and as it will be shown further, it is also absent in drosophila genome) in the genomes of arabidopsis and round worm is sufficiently noticeable and amounts ~1%.

If not take into account some particularities for chromosomes II and III, histograms of *A. thaliana* genome can be considered as similar (see fig. 7). Furthermore, the histograms for *C. elegans* chromosomes (see fig. 8) can be considered as practically identical.

Histograms based on the results of structure analysis of periodicity regions found out on the chromosomes of fruit fly *D. melanogaster* are shown in fig. 9. One can see that among

tandem repeats in *D. melanogaster* genome highly eroded (~1.5% – 2%) and eroded (~0.5% – 1%) micro-satellites are dominating. Similarity of the histograms in fig. 9 for the chromosomes 2L, 3L, 2R and 3R should be also noted.

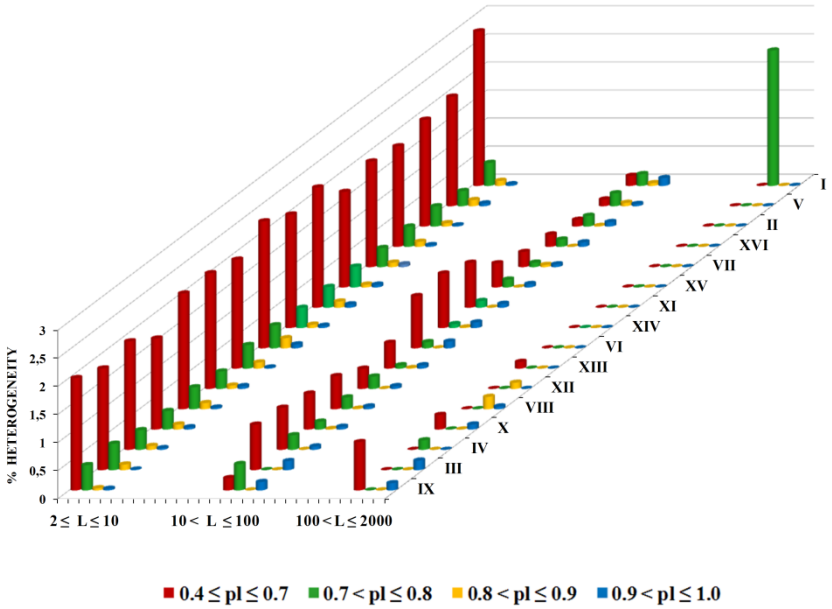

Figure 6. Histograms of structural content of the latent periodicity regions in *S. cerevisiae* genome (chromosomes I – XVI).

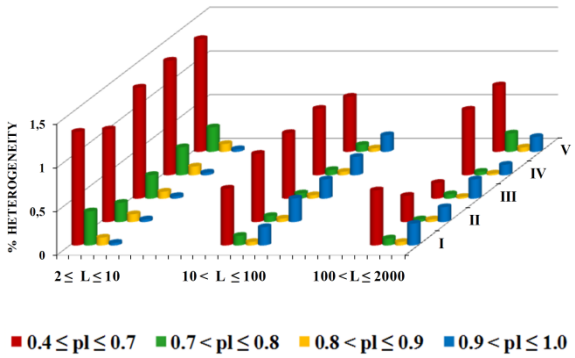

Figure 7. Histograms of structural content of the latent periodicity regions in *A. thaliana* genome (chromosomes I – V).

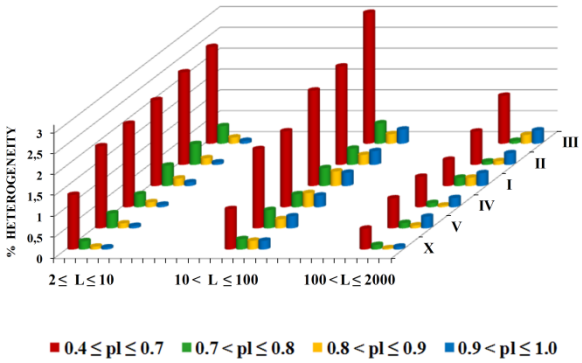

Figure 8. Histograms of structural content of the latent periodicity regions in *C. elegans* genome (chromosomes I – V, X).

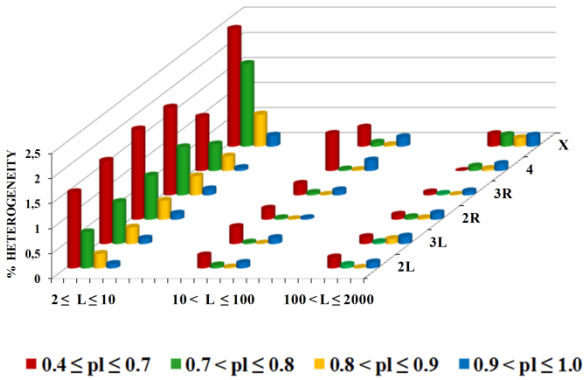

Figure 9. Histograms of structural content of the latent periodicity regions in *D. melanogaster* genome (chromosomes2L, 3L, 2R, 3R, 4, X).

Though, similarity of the histograms presenting structural (qualitative) content of the periodicity regions for different chromosomes of model organisms analyzed in the work, cannot be evidence for single evolutionary origin of such the chromosomes, nevertheless, it allows supposing existence of similar mechanisms of evolutionary pression and divergence influenced by which the chromosomes have been formed. Besides, as it follows from the analysis of the figures 6 – 9, in each genome one or even few characteristic dominating periodicity types may be discriminated, for example, highly eroded micro-satellites in *S. cerevisiae* genome. As can be noted for *A. thaliana* and *C. elegans*, their genomes have similar percentages of characteristic periodicity types. It can be supposed that significant percent (~1.5%) of mini- and mega-satellites is the consequence of active recombination processes [3] in the genomes of arabidopsis and nematode. Micro-satellites domination in yeast genome is possibly due to large number of genome DNA replication during yeast propagation and, consequently, due to high frequency of "slippage" mistake [2] leading to elongation of the micro-satellite regions. Understanding of dominating periodicities' functional significance is subject of possible future investigations.

Revealing of the latent periodicity in functional genome regions

Following link at Sequence Viewer graphical interface (http://www.ncbi.nlm.nih.gov/projects/sviewer/), for each heterogeneity region in the HeteroGenome database information about the region intersection with annotated sequences of genome analyzed may be obtained.

Tables 3 and 4 present general view of distribution of the latent periodicity group representers over annotated (functional) and not annotated (unassigned) DNA sequences in the GenBank.

To estimate distribution of the HeteroGenome groups over annotated genome regions, rather nonstrict criterium has been used. If area of group and functional region intersection covered no less than 50% of the smaller sequence, in such case group considered was assigned to the region. Besides, in estimating groups' distribution over annotated regions alternative splicing was not taken into account, i.e. only single mRNA and the one coding sequence (CDS – Coding DNA Sequence) were considered.

As one can see from Tab. 3, according to chosen criterium in the genomes of *S. cerevisiae*, *A. thaliana*, *C. elegans* and *D. melanogaster*, correspondingly, 80%, 62%, 65% and 67% of the HeteroGenome groups are allocated in the genes. For the same list of organisms 18%, 37.4%, 35% and 31.4% groups are allocated in not annotated (unassigned) regions of their genomes. Generally, distribution over the rest functional regions has accidental and negligible character. Though, it should be noted that 2.6% groups of *D. melanogaster* genome are placed in the regions of various repeats.

Table 3. Quantitative distribution of the HeteroGenome group representers over functional genome regions annotated in the GenBank

GenBank feature	*S. cerevisiae*[*]	*A. thaliana*	*C. elegans*	*D. melanogaster*
gene	3276	21598	25551	48657
mRNA	–	15337	12667	23879
CDS	3269	13370	12046	19463
intron	6	3105	14145	28112
exon	–	–	–	21
STS	–	141	–	150
rep_origin	15	–	–	–
repeat_region	95	–	–	1951
unassigned	738	12935	13773	22851
Total number of HeteroGenome groups	4094	34566	39329	72772

[*]Characteristics mRNA is absent in annotation of *S. cerevisiae* genome.

In locating group in gene there is probability to attribute the group both to exon and intron while the group is crossing their boarder. To avoid such double attribution additional analysis has been done taking into consideration the quantities of nucleotides in group which belong to exon and intron, correspondingly. Table 4 shows the results of this analysis in comparison with total number of the nucleotides in functional regions considered in the genome.

Table 4. Total quantity of the nucleotides from the HeteroGenome groups contained in the genes, exons and introns (Share of such the nucleotides in summarized length of the genes exons and introns of genome, correspondingly, is shown in brackets.)

Organism	Number of nucleotides from the HeteroGenome groups / number of corresponding functional nucleotides in the genome		
	genes	exons[*]	introns[*]
S. cerevisiae	354459 / 8829668, (4%)	352781 / 8737430, (4%)	448 / 64756, (0,7%)
A. thaliana	2037728 / 70751773, (3%)	1601059 / 40167753, (4%)	296614 / 19355644, (1,5%)
C. elegans	3816463 / 60377579, (6,3%)	1479515 / 27267246, (5,4%)	402554 / 32373603, (10%)
D. melanogaster	3786123 / 79344223, (4,8%)	3410523 / 28271547, (12%)	4419548 / 49121309, (9%)

[*]Data on the exons and introns have been obtained in accordance with indices of mRNA assembly ((join attribute line). For *S. cerevisiae* genome the indices of CDS assembly have been used.

Comparison of Tab. 4 and Tab. 2 shows that percentage of the nucleotides in groups attributed to the genes practically coincides with share of genome coverage by the regions of significant heterogeneity (latent periodicity). Distribution of the nucleotides among the exons and introns depends on organism. For example, percentage of total length for the latent periodicity regions in the exons of *D. melanogaster* (12%) is higher than in the introns (9%), despite summarized introns' length is nearly twice greater than summarized exons' length. And in contrast, having comparable total length for the exons and introns in *C. elegans* genome, share of periodicity regions in the introns is twice more than in the exons (10% and

5.4%, correspondingly). Consequently, direct dependence of the latent periodicity regions' length from the genome total length or from the length of functional regions is not observed.

Latent periodicity regions' density distribution over the chromosomes

Investigation of density distribution for the latent periodicity regions over the chromosomes was done for all organisms analyzed in the present work. So, each chromosome was divided into sequential fragments of equal length corresponding to 0.5% of total chromosome length. Fragment length is called step size of division. For each fragment summarized length (in number of nucleotides) of the latent periodicity regions located within the fragment boarders was determinated. Such length, being normed at total chromosome length and multiplied by 100%, was considered as a part of general share of the chromosome latent periodicity that is included in fragment considered. Summarizing over all fragments of division gives an estimate of general percentage of the latent periodicity regions on chromosome.

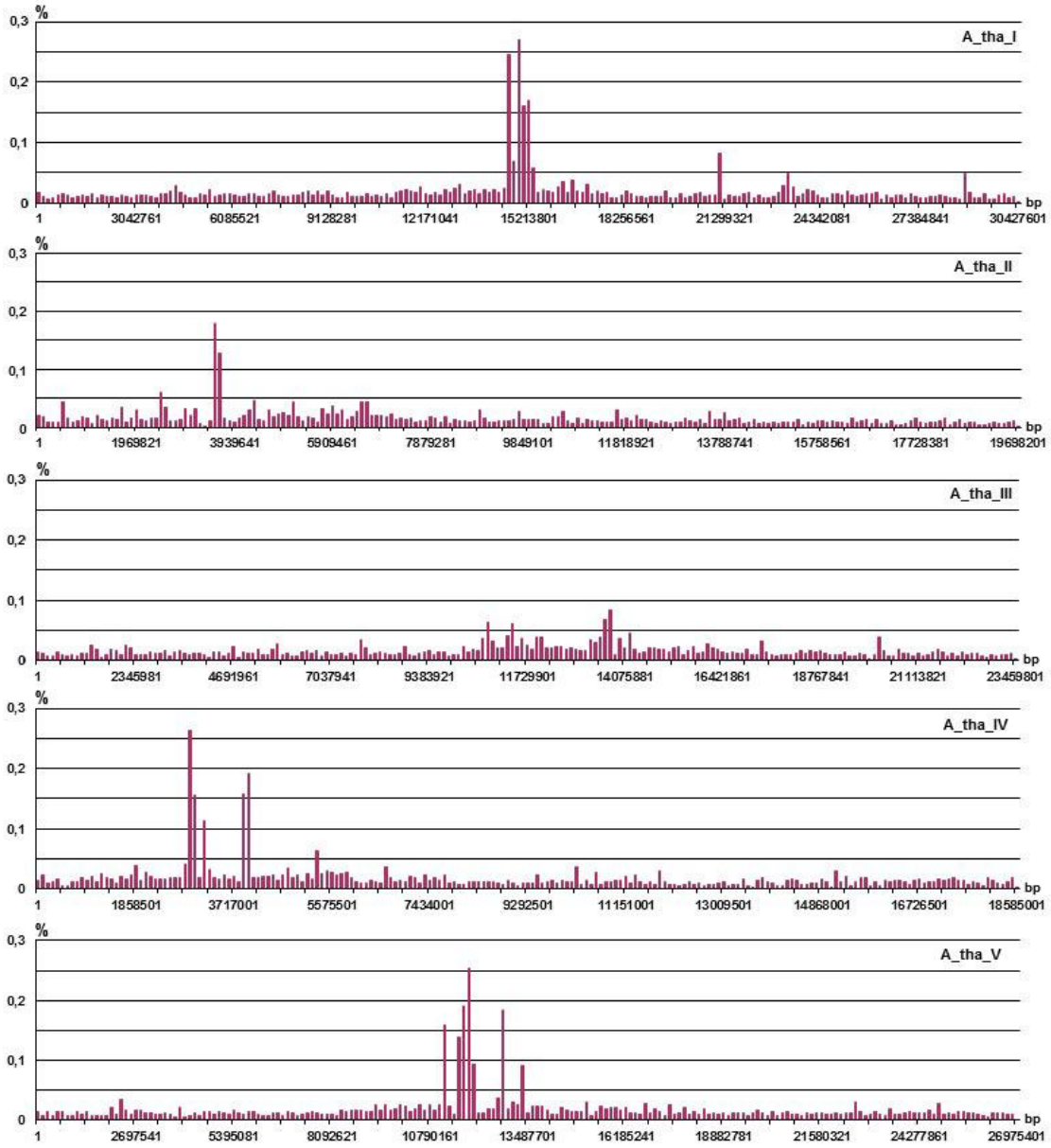

Figure. 10. Density distribution of the latent periodicity regions on the chromosomes of *A. thaliana*. High of histogram column shows percentage of local latent periodicity regions (relatively total chromosome length) revealed within the borders of corresponding histogram step.

While investigating density distribution of these regions only the group representer sequences, allowing nonredundant length estimation for chromosome coverage by the latent periodicity, were considered.

Histograms in figure 10 demonstrate density distribution of the latent periodicity regions for all chromosomes of the arabidopsis genome. Corresponding step size of division on chromosomes I – V are: 152138, 98491, 117299, 92925, 134877 nt. Results of the distributions over all chromosomes of the organisms considered are shown at page "Database Statistics" in the HeteroGenome database.

What is more, different distributions have been obtained over each chromosome for three classes of the latent periodicity, i.e. for micro- (period length $2 \le L \le 10$), mini- ($10 < L \le 100$) and mega- ($L > 100$) satellites. Example of such the distributions over chromosome I from *A. thaliana* genome is shown in the figure 11. As one can see from the histograms in the fig. 10 and fig. 11, density distribution of the latent periodicity regions unambiguously characterizes each chromosome from the genomes of organisms considered. Such distribution can be considered as a kind of DNA-fingerprint or individual bar-code for each chromosome in the genomes of various organisms.

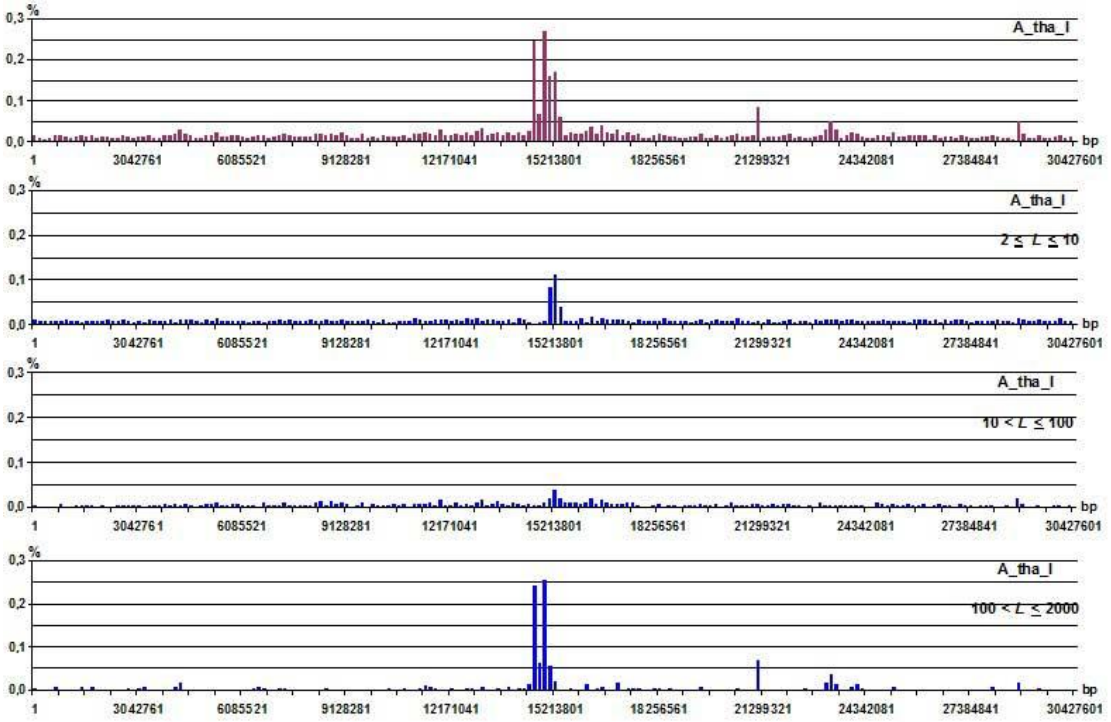

Figure. 11. Decomposition of general density distribution of the latent periodicity regions revealed on chromosome I from *A. thaliana* genome (top histogram) into three kinds of periodicity (micro-, mini- and mega-satellites).

Repertory of period lengths in the genome

All estimates obtained for the latent periodicity pattern lengths have been analyzed from point of view how frequently a certain period length may be encountered on each chromosome for every organism. Figures 12 and 13 demonstrate qualitative picture of frequencies for different periodicity pattern lengths. Periodicity pattern lengths ranging from 2 to 120 nt are marked along horizontal axis, while natural logarithm of number of the latent periodicity regions with a certain pattern length is shown along vertical axis. Graphics in the fig. 12, corresponding to *S. cerevisiae* and *D. melanogaster* genomes, show that "repertory" of the period lengths (i.e. series of the values for the latent periodicity pattern lengths) is practically the same for all chromosomes of any selected genome. However, in spite of "repertory" specifics for each genome common traits can be outlined in some cases. For

example, for chromosomes of *S. cerevisiae* and *D. melanogaster*all characteristics period lengths are, practically, multiplied by three. It is due to that cross area of the latent periodicity regions with genes encoding proteins in the genomes of *S. cerevisiae* and *D. melanogaster* correspondingly constitutes 87% and 66% of the revealed regions' total length. Peaks in the graphics point at characteristic pattern lengths among that the lengths of 3, 6, 12, 15, 18, 21, 24, 27, 33, 36, 45, 57 are common for both organisms.

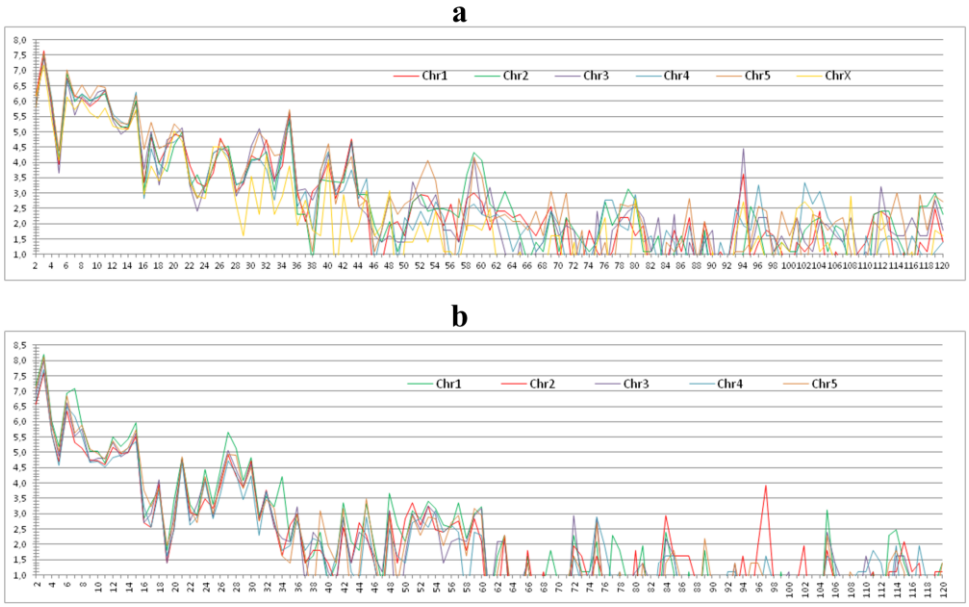

Figure 12. Graphics presenting repertory of the latent periodicity pattern lengths in the genomes of (**a**) *C. elegans* and (**b**) *A. thaliana*. Graphics corresponding to different chromosomes are drawn by alternative colors.

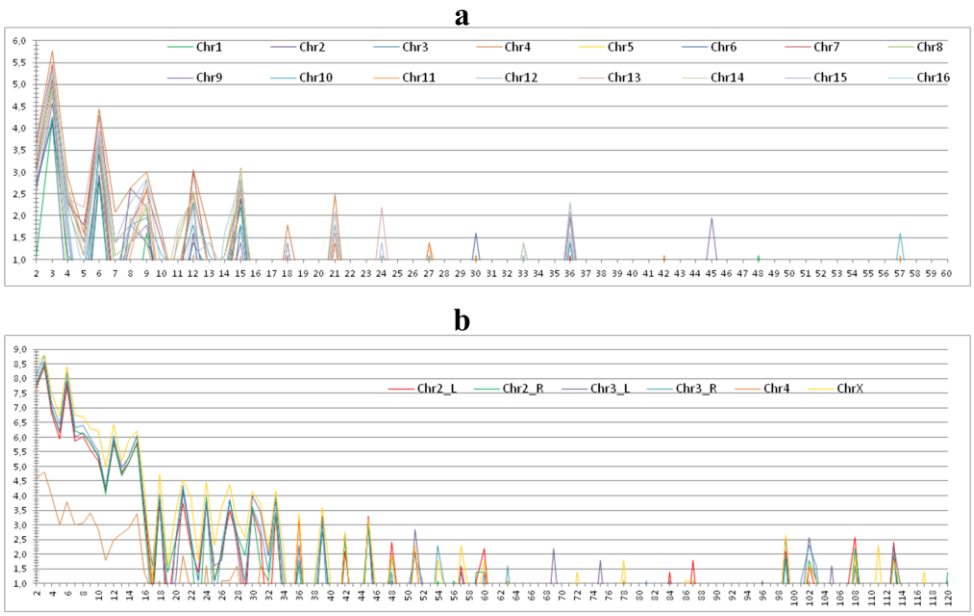

Figure 13. Graphics presenting repertory of the latent periodicity pattern lengths in the genomes of (**a**) *S. cerevisiae* and (**b**) *D. melanogaster*. Graphics corresponding to different chromosomes are drawn by alternative colors.

Graphics in the fig. 13, corresponding to the genomes of *C. elegans* and *A. thaliana* demonstrate the more reach "repertory" of characteristic period lengths for these organisms, where additional values exist which are greater or less the characteristics values (multiplied by three) on one unity. This may be due to the two factors. First, the less part of revealed

coverage by the latent periodicity regions in these genomes is situated in the genes encoding proteins (55% and 38% for *C. elegans* and *A. thaliana*, correspondingly). Secondly, it can be supposed that mutation processes in these genomes occur more extensively. In comparing the fig. 6 and fig. 9 (*S. cerevisiae* and *D. melanogaster*) with the fig. 7 and fig. 8 (*C. elegans* and *A. thaliana*), one can see that micro-satellites, formed mainly in the result of DNA polymerase slippage during replication, dominate in the genomes of the first two organisms. Besides micro-satellites significant part of mini- and mega-satellites, propagating in consequence of various recombination processes and genome duplications, are revealed in *C. elegans* and *A. thaliana* genomes.

CONCLUSIONS

Reliable data about the tandem repeats, including highly eroded repeats, selected for the first release of the HeteroGenome database (http://www.jcbi.ru/lp_baze/), have been obtained in the result of spectral-statistical approach employment for revealing heterogeneity (latent periodicity) regions in the genomes of model organisms *S. cerevisiae*, *A. thaliana*, *C. elegans* and *D. melanogaster*.

Specially elaborated two-level structure of logical record in the database allowed presenting data as not intersected regions of the latent periodicity on the chromosomes (nonredundant data representation) and pointing at more conservative areas of periodic structure of such the regions either.

Owing to user friendly interface and possibility of additional data analysis the HeteroGenome database can be useful for molecular genetic research of the model organisms and further investigation of the latent periodicity phenomenon in DNA sequences.

Conclusion that latent periodicity regions constitute ~10% in the genomes of various organisms can be done in the result of data obtained analysis for the genomes of *S. cerevisiae*, *A. thaliana*, *C. elegance* and *D. melanogaster*. Highly eroded micro-satellites (with period length ~10 nt) dominate in all model organisms mentioned above and constitute ~2% of their genome length. It was shown in the work that characteristic quantitative and qualitative contents of the latent periodicity regions are revealed for each genome. For example, in the genomes of yeast *S. cerevisiae* (excluding the chromosomes I and IX) and fruit-fly *D. melanogaster* mega-satellite repeats (with period length of more than 100 nt) are practically absent. Otherwise, percentage of the mega-satellite repeats in the genomes of cress-tale *A. thaliana* and round worm *C. elegance* is sufficiently noticed and constitutes ~1%.

Analysis of the HeteroGenome base data distribution over functional (annotated in the GenBank) and not functional (unassigned) shows that more than one-half of the latent periodicity regions are found out in the genes encoding proteins.

Obtained density distributions of the latent periodicity regions on the chromosomes demonstrate their chromosome specificity. It seems likely, that picture of the periodicity regions' distribution is unique characteristics or even an identifier of each chromosome and, eventually, it probably reflects the positioning of transcriptionally active DNA sequences.

Repertories of the latent periodicity pattern lengths in the genomes of baker yeast, cress-tale, round worm and fruit-fly have been analyzed. It was shown that the genome of any organism has its' specific repertory and for the same genome repertories of all chromosomes are practically coincided. Some organisms have pure repertories with certainly defined characteristic period lengths as, for example, yeast (*S. cerevisiae*) and fruit-fly (*D. melanogaster*). For other two organisms a large variety of period lengths is observed. As may be supposed it is a consequence of actively occurring processes of recombination and duplication in the genome.

The work has been partially supported by grant № 12-07-00530 of Russian foundation for basic research (RFBR).

REFERENCES

1. Richard G.F., Kerrest A., Dujon B. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiol. Mol. Biol. Rev.* 2008. V. 72. P. 686–727.
2. Kelkar Y.D., Strubczewski N., Hile S.E., Chiaromonte F., Eckert K.A., Makova K.D. What is a microsatellite: a computational and experimental definition based upon repeat mutational behavior at A/T and GT/AC repeats. *Genome Biol. Evol.* 2010. V. 2. P. 620–635.
3. Ellegren H. Microsatellites: simple sequences with complex evolution. *Nat. Rev. Genet.* 2004. V. 5. P. 435–445.
4. Welch J.W., Maloney D.H., Fogel S. Unequal crossing-over and gene conversion at the amplified CUP1 locus of yeast. *Mol. Gen. Genet.* 1990. V. 222. P. 304–310.
5. Tyler-Smith,C. and Willard,H.F. Mammalian chromosome structure. *Curr. Opin. Genet. Dev.* 1993. V. 3. P. 390–397.
6. Hewett D.R., Handt O., Hobson L., Mangelsdorf M., Eyre H.J., Baker E., Sutherland G.R., Schuffenhauer S., Mao J.I., Richards R.I. FRA10B structure reveals common elements in repeat expansion and chromosomal fragile site genesis. *Mol. Cell.* 1998. V. 1**.** P. 773–781.
7. Yu S., Mangelsdorf M., Hewett D., Hobson L., Baker E., Eyre H.J., Lapsys N., Le Paslier D., Doggett N.A., Sutherland G.R., Richards R.I. Human chromosomal fragile site FRA16B is an amplified AT-rich minisatellite repeat. *Cell*. 1997. V. 88. P. 367–374.
8. Fu Y.H., Kuhl D.P., Pizzuti A., Pieretti M., Sutcliffe J.S., Richards S., Verkerk A.J., Holden J.J., Fenwick R.G. Jr, Warren S.T., et al. Variation of the CGG repeat at the fragile X site results in genetic instability: resolution of the Sherman paradox. *Cell*. 1991. V. 67. P. 1047–1058.
9. Liquori C.L., Ricker K., Moseley M.L., Jacobsen J.F., Kress W., Naylor S.L., Day J.W., Ranum L.P. Myotonic dystrophy type 2 caused by a CCTG expansion in intron 1 of ZNF9. *Science*. 2001. V. 293. P. 864–867.
10. Matsuura T., Fang P., Pearson C.E., Jayakar P., Ashizawa T., Roa B.B., Nelson D.L. Interruptions in the expanded ATTCT repeat of spinocerebellar ataxia type 10: repeat purity as a disease modifier? *Am. J. Hum. Genet.* 2006. V. 78. P. 125–129.
11. Lalioti M.D., Scott H.S., Buresi C., Rossier C., Bottani A., Morris M.A., Malafosse A., Antonarakis S.E. Dodecamer repeat expansion in cystatin B gene in progressive myoclonus epilepsy. *Nature*. 1997. V. 386. P. 847–851.
12. Martin P., Makepeace K., Hill S.A., Hood D.W., Moxon E.R. Microsatellite instability regulates transcription factor binding and gene expression. *Proc. Natl. Acad. Sci. USA*. 2005. V. 102. P. 3800–3804.
13. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*. 1999. V. 27. P. 573–580.
14. Reneker J., Shyu C.R., Zeng P., Polacco J.C., Gassmann W. ACMES: fast multiple-genome searches for short repeat sequences with concurrent cross-species information retrieval. *Nucleic Acids Res*. 2004. V. 32. P. W649–W653.
15. Roset R., Subirana J.A., Messeguer X. MREPATT: detection and analysis of exact consecutive repeats in genomic sequences. *Bioinformatics*. 2003. V. 19. P. 2475–2476.
16. Parisi V., Fonzo V.D., Aluffi-Pentini F. STRING: finding tandem repeats in DNA sequences. *Bioinformatics*. 2003. V. 19. P. 1733–1738.
17. Kolpakov R., Kucherov G. Mreps: efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Res.* 2003. V. 31. P. 3672–3678.
18. Wexler Y., Yakhini Z., Kashi Y., Geiger D. Finding approximate tandem repeats in genomic sequences. *J. Comput. Biol.* 2005. V. 12. P. 928–942.

19. Boeva V., Regnier M., Papatsenko D., Makeev V. Short fuzzy tandem repeats in genomic sequences, identification, and possible role in regulation of gene expression. *Bioinformatics*. 2006. V. 22. P. 676–684.
20. Mudunuri S.B., Nagarajaram H.A. IMEx: imperfect microsatellite extractor. *Bioinformatics*. 2007. V. 23. P. 1181–1187.
21. Pellegrini M., Renda M.E., Vecchio A. TRStalker: an efficient heuristic for finding fuzzy tandem repeats. *Bioinformatics*. 2010. V. 26. P. i358–i366.
22. Sokol D., Benson G., Tojeira J. Tandem repeats over the edit distance. *Bioinformatics*. 2007. V. 23. P. e30–e35.
23. Sokol D., Atagun F. TRedD – A database for tandem repeats over the edit distance. *Database*. 2010. Article ID baq003.
24. Gelfand Y., Rodriguez A., Benson G. TRDB – the Tandem Repeats Database. *Nucleic Acids Res.* 2007. V. 35. P. 80–87.
25. Boby T., Patch A., Aves S. TRbase: a database relating tandem repeats to disease genes for the human genome. *Bioinformatics*. 2005. V. 21. P. 860–921.
26. Chaley M.B., Nazipova N.N., Kutyrkin V.A. Statistical methods for detecting latent periodicity patterns in biological sequences: the case of small-size samples. *Pattern Recogn. Image Anal.* 2009. V. 19. P. 358–367.
27. Chaley M.B., Nazipova N.N., Kutyrkin V.A. Joint use of different homogeneity testing criteria for latent periodicity revelation in biological sequences. *Math. Biol. Bioinf.* 2007. V. 2(1). P. 20–35. doi: 10.17537/2007.2.20. published in Russian.
28. Chaley M., Kutyrkin V. Model of perfect tandem repeat with random pattern and empirical homogeneity testing poly-criteria for latent periodicity revelation in biological sequences. *Math. Biosci.* 2008. V. 211. P. 186–204.
29. Fields S., Johnston M. Cell biology. Whither model organism research? *Science*. 2005. V. 307. P. 1885–1886.
30. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*. 2001. V. 409. P. 860–921.

Received 23.07.2018, published 09.08.2018.