Mathematical Biology and Bioinformatics. 2013. V. 8. № 2. P. t27–t33.
URL: http://www.matbio.org/2013/Korotkov_8_t27.pdf.

Original Russian text
Korotkov E., Shelenkov A., Korotkova M., 2013 published in Matematicheskaya biologiya i bioinformatika, 2013. V. 8. № 2. P. 529–536. URL: http://www.matbio.org/2013/Korotkov_8_529.pdf.

========== **COMMENTS** ==========

Towards the identification of the latent periodicity in DNA sequences

Korotkov E., Shelenkov A., Korotkova M.

Bioengineering Centre, Russian Academy of Sciences, Moscow, Russia, 117312

Abstract. In this paper we compared the information decomposition (ID) method and the spectral-statistical (SS) approach. We showed that the SS approach does not take into account the effect of small samples, and it does erroneous search of statistically significant period in the DNA sequence. Detection of the "profile periodicity" by SS approach depends solely on the ratio of the lengths of the latent periods. The revealed drawbacks of the spectral-statistical approach show that to search for DNA regions with latent periodicity it is more consistent from mathematical point of view to use *Z*-statistic and the method of information decomposition.

Keywords: *latent periodicity, information decomposition, spectral-statistical approach, genes, triplet periodicity, profile periodicity.*

1. INTRODUCTION

In the publication [1] a comparison of the spectral-statistical (SS) approach and the information decomposition (ID) method [2] is made. Also the SS approach is used to analyze the sequences possessing latent periodicity in presence of insertions and deletions which were published in publication [3]. We think that the comparison of these two methods developed to perform the search for the latent periodicity in nucleotide sequences was made with errors and that the SS approach [1] itself has some serious drawbacks. In this paper we consider the errors made during comparison of the SS approach and the ID method, and the drawbacks of the spectral-statistical approach. Below we consider the drawbacks of the paper [1] and of some other papers published by the same authors.

2. THE SEARCH FOR LATENT PERIODICITY IN THE FRAGMENTS OF SEQUENCES WITH GENBANK IDS AF453480 AND CO11168X1

In the publication [1], pp. 510–511, the authors tried to compare the spectral-statistical approach developed by them with the method we used earlier in the papers [2–3]. However this comparison was made by the authors with a gross blunder. The aim of the paper [3] was to demonstrate the existence of nucleotide periodicity which could be revealed only in presence of insertions and deletions of nucleotides. By way of example, several such sequences were shown, inter alia, the region from 4166th to 4368th nucleotides of the sequence with Genbank identifier AF453480, and the region from 176412th to 176535th nucleotides of the sequence with Genbank identifier CO11168X1. The latent periodicity with a period length $\lambda = 2$ was revealed in these sequences only in presence of some nucleotide deletions and insertions. The alignments of these two sequences with the corresponding periodic consensus sequences were shown in the Table 4 of our paper [3]. However the authors of the paper [1] have applied their approach to these two sequences without using the

alignment made by us. We recalculated the results from the Fig. 5 (a and b) of the paper [1] without making the alignment, and the results of this recalculation are shown in the Fig. 1,A and Fig. 2,A. It is easy to see that the results are generally the same as in the Fig. 5 of the paper [1]. However, if we use the alignments made by us in the paper [3], the results are completely different and one can see the period with a length equal to 2 nucleotides. These results are shown below in the Fig. 1,B for the sequence AF453480 and in the Fig. 2,B for the sequence CO11168X1.

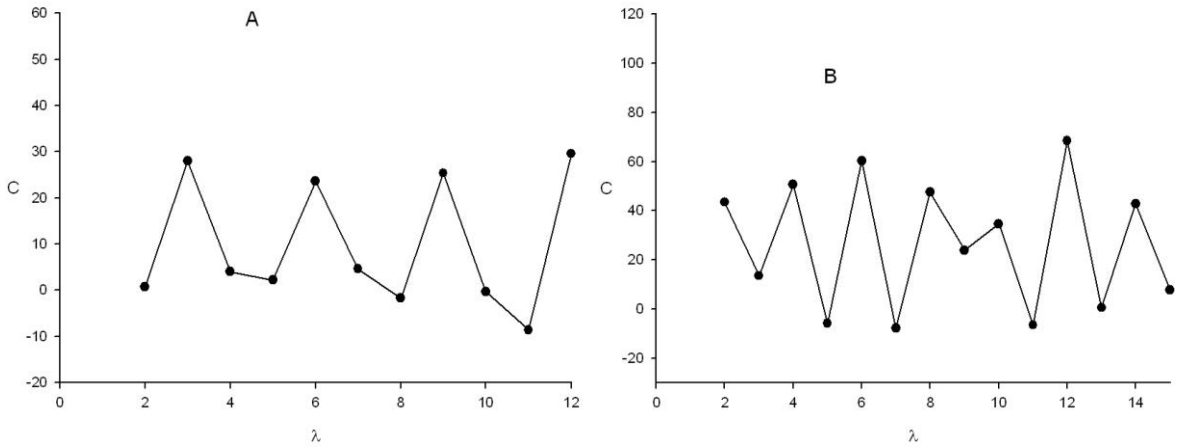

Fig. 1. Characteristic spectra for the region of the sequence with Genbank ids AF453480 from 4166th to 4368th nucleotides built without taking into account the alignment with consensus (A) and with using the alignment (B) made in the paper [3]. Here λ is a period length.

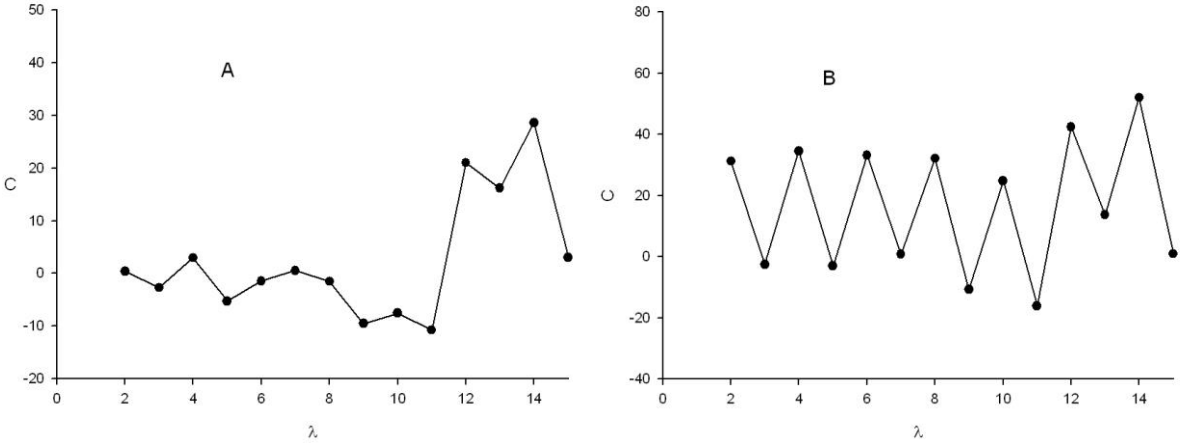

Fig. 2. Characteristic spectra for the region of the sequence with Genbank ids XO11168X1 from 176412th to 176535th nucleotides built without taking into account the alignment with consensus (A) and with using the alignment (B) made in the paper [3]. Here λ is a period length.

The graphs show that the comparison with our results in the publication [1] was made with a gross blunder since it was made without taking into account the alignment made by us (insertions and deletions of the nucleotides). So the conclusion in the paper [1] on the page 511, namely, «So the results of the latent periodicity estimation obtained in the paper [4] could appear inaccurate without their verification» is erroneous. Apparently, the authors of the publication [1] have not studied our paper [3] carefully and thus they were not able to make the accurate comparison of the results.

3. THE SMALL SAMPLE PROBLEM IN THE SPECTRAL-STATISTICAL APPROACH

In the paper [1] χ^2 distribution is used to estimate the statistical significance of the revealed periodicities. As it can be seen from this work of the authors and their previous publication [4], it was performed the analysis of short-length sequences in which just a few periods could possibly reside in a latent periodicity region. In this case the evaluation of the formulae (2) and (4) in the publication [1] would be performed for N periods, where N is small. The length of the sequence being analyzed equals to $N\lambda$, where λ is a period length. If the number of periods is small, namely, lies within the range from 2 to about 20 periods, then all statistical significance estimations made in the publication [1] become very inaccurate. This range is related to so-called "small sample" for which the theoretical estimates could be erroneous. This fact is well-known to all mathematicians working in the field of statistical analysis of the data. For example, this is described in the book [5].

For the purpose of illustration, in this work we estimated the small sample influence on the probability α that is used in the formula (4) of the publication [1] for the period length equal to $\lambda = 32$. The probability α in the papers [1, 4] was chosen to be equal to 0.05, and this value was used to calculate, for each period length λ, such a value $\chi^2_{0.05}$ that the probability would be $P(x \geq \chi^2_{0.05}) = 0.05$ for x distributed according to χ^2 with $R(\lambda-1)$ degrees of freedom. Here R is a number of different DNA bases used and equals to 4. In the Fig. 3 the dependence of the probability value α determined using the Monte-Carlo method on the number of periods N is shown. It can be seen that the deviation of α values obtained for the case of small sample from the expected values of α for a big sample can be five-fold. These deviations will be larger for the smaller probability values α taken, and such deviations also depend on nucleotide frequencies in the latent periodicity region revealed. Thus the statistical significance estimates made in the publication [1], as well as in previous publication [4], are inaccurate, allowing the error to be several hundred percent in some cases. Such errors lead to distortion of the spectra of "profile periodicity" built according to the formulae (2) and (4) from the publication [1] and to erroneous conclusions both in this publication and in earlier publications of these authors.

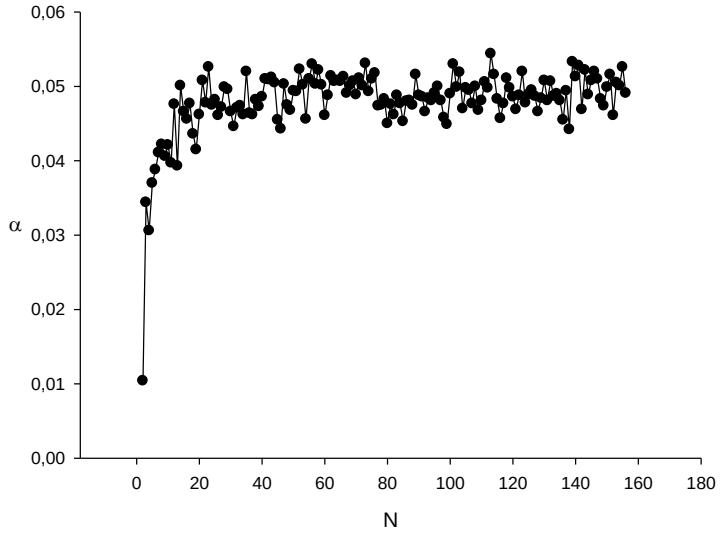

Fig. 3. The dependence of the probability α, that is used in the formulae (2) and (4) of the publication [1], on the number of periods N is shown. The calculation was made for the period length λ equal to 32 nucleotides and for uniform nucleotide frequencies.

4. SEARCH FOR THE PERIOD WITH THE GREATEST STATISTICAL SIGNIFICANCE IN SPECTRAL-STATISTICAL APPROACH

In the publication [1] a characteristic spectrum $C(\lambda)$ (formula (3) of the publication [1]) and a spectrum $D_L(\lambda)$ (formulae (2) and (4) of the publication [1]) are used to determine the latent period having maximal significance. The spectrum $C(\lambda)$ and the spectrum $D_1(\lambda)$ are completely similar to the spectra $2I$ in dependence on λ that were obtained by us in papers [6, 7]. The difference is that the spectrum $C(\lambda)$ can be obtained from the spectrum $2I(\lambda)$ by subtracting the mean value $E(2I(\lambda))$ for each λ, and the spectrum $D_1(\lambda)$ can be obtained by dividing for each λ the values of $2I(\lambda)$ on the values of $2I_{0.05}(\lambda)$ for which $P(2I > 2I_{0.05}(\lambda)) = 0.05$. The spectrum $C(\lambda)$ is used in the paper [1], pages 505–506, to search for the first test period with a maximal value of $C(\lambda)$. The authors of [1] in the bottom of the page 506 reported «the first test period L with clear-cut maximal value of a spectrum C serves as an estimate for the latent period in a string *str*.” Later on the same page they reported: «These spectra, according to the rule formulated, allow to obtain the right estimates of the latent periods for the tandem repeats considered”. We consider these statements to be erroneous in a general case. By "general case" we mean the case in which a DNA sequence being analyzed may contain one latent period or several such periods. Our statement is based on the fact that we cannot compare $C(\lambda)$ for different values of λ since when $C(\lambda_1) = C(\lambda_2) = C_0$, these values have totally different probabilities $P(C(\lambda_1) \geq C_0)$ and $P(C(\lambda_2) \geq C_0)$ according to χ^2 distribution used by the authors of [1]. We have calculated, by way of example, the probabilities $P1$ of that $C(\lambda) \geq 30.0$ for the values of λ in the interval from 2 to 60. The results are shown in the Fig.4. In this figure one can see that these probabilities differ dramatically. If we take into account the form of the function shown in the Fig.4, it can be seen that longer periods having smaller statistical significance will always have preference upon shorter periods. This means that the true latent period length and its statistical significance could be determined incorrectly based on the spectrum $C(\lambda)$.

Also, the latent period having maximal significance cannot be determined based on the spectrum $D_1(\lambda)$ since when $D_1(\lambda_1) = D_1(\lambda_2) = D_0$, these values have totally different probabilities $P(D_1(\lambda_1) \geq D_0)$ and $P(D_1(\lambda_2) \geq D_0)$ according to χ^2 distribution used by the authors of [1] (Fig. 5).

In the paper [1] there is also an error in the Fig. 2,b. As it can be seen from the Fig. 4,b, the authors apply their rule that we cited in the previous section, and from all the periods visible in the Fig. 4,b they choose the one having the maximal value of $C(\lambda)$, i.e., they skip all values of λ up to $\lambda = 84$, for which the maximum of $C(\lambda)$ is observed. But if we apply this rule to the Fig. 2,b, we should acknowledge the presence of the period with a length equal to 24 nucleotides in the sequence analyzed, since it can be seen from this figure that $C(12) \approx 48.0$ and $C(24) \approx 60.0$, i.e., in any case $C(12) < C(24)$. So it remains a mystery why in this case the authors contradict their rule and why do they think that this rule allows to obtain "the accurate estimates of the latent periods". The authors of the paper [1] do not present any arguments or calculations to prove the correctness of their rule.

It should be noted that the situations in which the rule developed by the authors gives the erroneous period length will often occur while analyzing nucleotide sequences. We contemplate the situation in which the maximum is in $C(k\lambda)$, where k changes from 2 to the value for which $\lambda_{max} \geq k\lambda$ still holds. Here λ_{max} is a maximal period length in a spectrum $C(\lambda)$ for the sequence analyzed. The point is that in this case, as we reported earlier in the paper [7], $\chi^2(k\lambda) \geq \chi^2(\lambda)$ if we use the formula (1) of the publication [1], i.e. $\chi^2(\lambda)$ is completely included in $\chi^2(k\lambda)$. Thus if due to random factors the value $C(k\lambda) = C_k$ becomes greater than the value $C(\lambda) = C_0$, then the probability $P(C(k\lambda) > C_k)$ can become greater than the probability $P(C(\lambda) \geq C_0)$. But for the period with the greatest statistical significance this probability should be minimal. Practically, this will lead to erroneous determination of the period with the greatest statistical significance which authors determine in the publication [1]

and some their previous publications. We think that this error also occurs in the paper [8] published by the authors, in which a dendrogram showing separation of some coding regions into groups according to the revealed properties of their sequences is presented in the Fig. 5. In the publication [8] a separation of the coding sequences on the 4th level is made into "possessing 3-profile" and "not possessing 3-profile" ones. Totally on this level 12996 sequences were contained (73.6% of the total number of sequences). These sequences were divided into two groups. The sequences from the first group consisting of 10699 sequences (60.6% of the total number of source sequences) were referred to "possessing 3-profile", and the sequences from the second group containing 2297 sequences (13% of the total number of source sequences) were referred to as "not possessing 3-profile". Due to the facts mentioned above such a separation performed according to the method described in the papers [1, 8] will be incorrect since the period length having maximal statistical significance is determined incorrectly. In fact, this will lead to dramatic decrease of the number of DNA regions possessing periodicity with a length equal to 3 nucleotides and to corresponding increase of the number of periods with other lengths in the Fig. 5–7 of the publication [8].

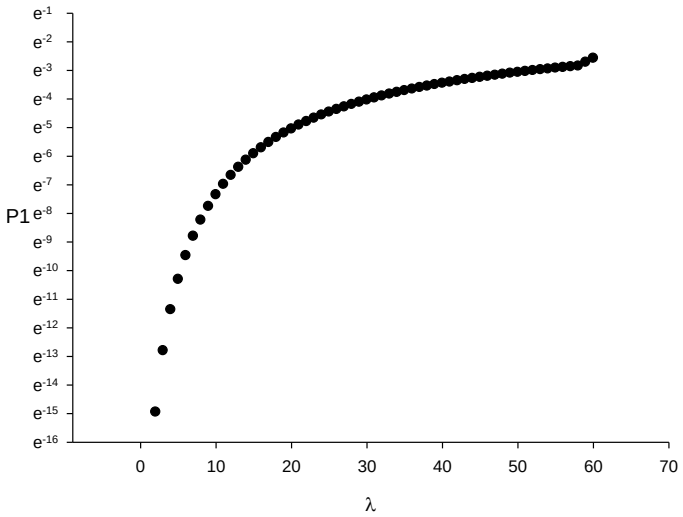

Fig. 4. A probability $P1$ of $C(\lambda)\geq 30.0$ is shown for the values of λ (period length) in a range from 2 to 60.

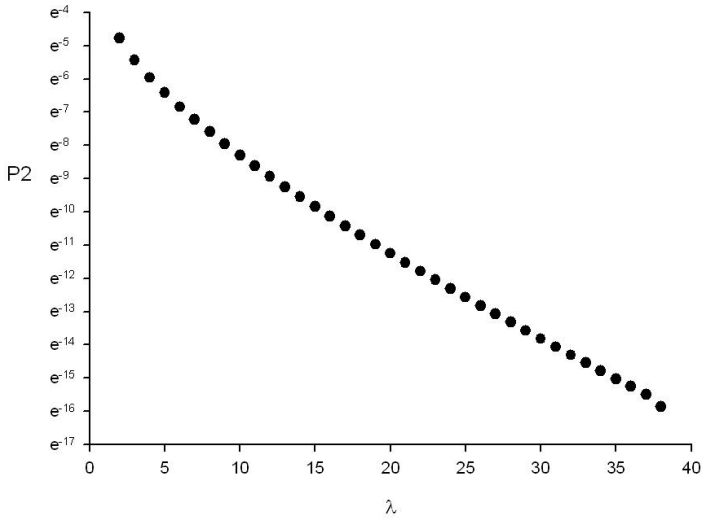

Fig. 5. A probability $P2$ of $D_1(\lambda) \geq D_0$ is shown for the values of λ (period length) in a range from 2 to 38 and for $D_0 = 1.5$.

5. THE CONCEPT OF "PROFILE PERIODICITY"

We also believe that the separation of the latent periodicity into profile and non-profile ones [1, 8] does not reflect any characteristic property of the latent periodicity, but rather depends only on the ratio of the lengths of the latent periods contained in the sequence analyzed. The point is that a nucleotide sequence may possess several latent periodicities with different period lengths. According to the data from the paper [7], the periods which lengths are coprime integers do not have influence on each other. These period lengths can "induce" the periodicity of all divisible lengths. This means that $\chi^2(\lambda)$ is completely included in $\chi^2(k\lambda)$ if we use the formula (1) from the publication [1], as we mentioned above in the section 4. If the period length is a composite number, then such an "inducing" occurs for all lengths equal to the prime numbers contained in this composite number and equal to the products of these prime numbers. This phenomenon can be clearly seen in the Fig. 4,b of the publication [1]. So if we have several periodicities of different lengths in a sequence and these lengths are prime numbers, then the profile period will always be the period with a length equal to their product. For this period length all values of D will always be less than 1. This is perfectly demonstrated in the Fig. 3 of the paper [4]. In that case we have two periodicities with lengths equal to 3 bases and 11 bases, respectively. Thus the profile period have a length of 33 bases. This length is less than λ_{max} (see section 4), so it is possible to reveal the "profile" periodicity. If the latent periods, for example, had lengths equal to 37 bases and 13 bases, respectively, and $\lambda_{max} = 300$, then in such a sequence the methods proposed in the papers [1, 4 ,5] would not be able to reveal the profile periodicity simply because $13 \times 37 > 300$. Therefore the presence or the absence of "profile periodicity" can serve only as a method of differentiation of latent periodicities in a sequence analyzed based on the product of period lengths for the periods contained in it, and this concept lacks any other meaning. This is why the classification of coding sequences from KEGG database into profile and non-profile ones made in Fig.5 in the paper [8] shows that only for 20.8% of genes the length of the "profile period" is greater than the length of the gene analyzed. This means that many "non-profile sequences" contain the periodicity with a length equal to 3 bases with a background of one or more longer periods. This is why we believe that the concept of "profile periodicity" introduced in the papers [1, 4] in not informative to make conclusions regarding the presence or the absence of latent periodicity in nucleotide sequences.

6. TWO LEVEL STRUCTURE OF LATENT PERIODICITY IN GENES

The two level coding structure that is mentioned in the paper [8], page 150 and in the publication [4] was noticed earlier by us in the paper [7], page 204. However in the papers [4, 8] a reference to the publication [7] is missing. We have also compared the effect described in the paper [7] with amplitude modulation of radio-frequency signals. The latent periodicity with a period length divisible by 3 somewhat modulates the periodicity of the length 3 in genes. Thus we think that the authors of the papers [4, 8] have not noticed this fact by themselves, but have rather copied the conclusions of the work [7] without making reference to the original source in this case.

7. THE ADVANTAGES OF THE INFORMATION DECOMPOSITION METHOD

Upon studying the disadvantages of the spectral-statistical approach described in sections 3–5 it becomes evident that the usage of Z statistic for searching the latent periodicity in nucleotide sequences (publications [2, 3, 6, 7]) is more reasonable than the usage of the spectra $C(\lambda)$, $D(\lambda)$ (formulae (3) and (4) of the publication [1]) and the usage of χ^2 criterion according to formula (1) from the publication [1]. First, the influence of small sample statistic decreases significantly. Second, using Z value it is possible to compare the statistical significance of the latent periodicity with different period lengths. If $Z(\lambda_1) = Z(\lambda_2) = Z_0$, then this means that the probabilities $P(Z(\lambda_1) \geq Z_0)$ and $P(Z(\lambda_2) \geq Z_0)$ will also be equal to each

other. This allows us to find the period with the greatest statistical significance (having the maximal Z value) based on the spectrum $Z(\lambda)$. This is not the case when using the statistics $C(\lambda)$, $D(\lambda)$ or with direct usage of χ^2 (publication [1], formulae (3), (4) and (1), respectively).

8. MISSING DATA FOR THE TYPE I AND TYPE II ERRORS

In the publication [1] a dendrogram building for the coding sequences is performed based on the spectral-statistical approach. Each step of such a classification requires the determination of FDR (false discovery rate) and the number of type I and type II errors (false positives and false negatives). However, FDR and the number of type I and type II errors were not determined in this paper. Therefore the separation of the gene coding regions into uniform and non-uniform ones, into profile and non-profile ones and into 3-regular and 3-nonregular ones seems not to be statistically significant (Fig. 5 from the paper [1]). Also in this classification an implicit dependences on the length distribution of coding sequences and on the maximal period length λ_{max} in profile spectrum are present, which should not be the case for a mathematically sound classification. Such dependences make the dendrogram built biologically irrelevant. The facts mentioned above do not allow to consider the coding sequences' classes introduced in this paper as mathematically or biologically reasonable.

In the publication [9] a search for the latent periodicity regions in DNA sequences from various chromosomes was performed. This paper also lacks the determination of FDR (false discovery rate) and type I and type II errors. In addition, the database created inherits all the disadvantages of the "spectral-statistical approach" (see sections 3–5 above). Therefore the database seems to contain incorrect information.

REFERENCES

1. Chaley M.B., Kutyrkin V.A. Recognition of latent periodicity in DNA sequences. *Mathematical Biology and Bioinformatics.* 2013. V. 8. № 2. P. 502–512.
2. Korotkov E.V., Korotkova M.A., Kudryashov N.A. Information decomposition method for analysis of symbolical sequences. *Physical Letters A*. 2003. V. 312. P. 198–210.
3. Shelenkov A., Skryabin K., Korotkov E. Search and classification of potential minisatellite sequences from bacterial genomes. *DNA Res*. 2006. V. 13. P. 89–102.
4. Chaley M., Kutyrkin V. Profile-statistical Periodicity of DNA. Coding Regions. *DNA Res.* 2011. V. 18. P. 353–362.
5. Suchoruchenkov B.I. *Analysis of a small sample. Applied statistical methods*. Moscow, 2010. (in Russ.)
6. Korotkov E.V., Korotkova M.A. DNA regions with latent periodicity in some human clones. *DNA Sequence.* 1995. V. 5. P. 353–358.
7. Korotkov E.V., Korotkova M.A., Tulko J.S. Latent sequence periodicity of some oncogenes and DNA-binding protein genes. *CABIOS*. 1997. V. 13. P. 37–44.
8. Chaley M.B., Kutyrkin V.A. Structural differences coding and noncoding regions of the DNA sequences of the human genome. *Bulletin of MGTU of N.E. Bauman, Ser. Natural sciences*. 2012. № 3. P. 146–157.
9. Chaley M.B., Kutyrkin V.A., Tuylbasheva G.E., Teplukhina E.I., Nazipova N.N. Investigation of Latent Periodicity Phenomenon in the Genomes of Eukaryotic Organisms. *Mathematical Biology and Bioinformatics.* 2013. V. 8. № 2. P. 480–501.

Received October 14, 2013.
Published December 11, 2013.