Journal of Siberian Federal University. Biology 3 (2015 8) 278-286
~ ~ ~

УДК 57:015 + 573.2

# Symmetry of Siberian Larch Transcriptome

**Michael G. Sadovsky[a]*, Vladislav V. Birukov[b], Yuliya A. Putintseva[b], Nataliya V. Oreshkova[b,c], Eugene A. Vaganov[b] and Konstantin V. Krutovsky[b,d,e,f]**

*[a]Institute of Computational Modelling SB RAS*
*50/44 Akademgorodok, Krasnoyarsk, 660036, Russia*
*[b]Siberian Federal University*
*Genome Research and Education Centre*
*50a/2 Akademgorodok, Krasnoyarsk, 660036, Russia*
*[c]V. N. Sukachev Institute of Forest SB RAS*
*50/28 Akademgorodok, Krasnoyarsk, 660036, Russia*
*[d]Georg-August-University of Göttingen*
*2 Büsgenweg, Göttingen, D-37077, Germany*
*[e]N. I. Vavilov Institute of General Genetics RAS*
*3 Gubkin Str., Moscow, 119333, Russia*
*[f]Texas A&M University*
*HFSB 305, 2138 TAMU, College Station, Texas, 77843, USA*

Received 10.03.2015, received in revised form 14.04.2015, accepted 14.07.2015

*The paper presents a novel approach to infer a structuredness in a set of symbol sequences such as transcriptome nucleotide sequences. A distribution pattern of triplet frequencies in the Siberian larch (Larix sibirica Ledeb.) transcriptome sequences was investigated in the presented study. It was found that the larch transcriptome demonstrates a number of unexpected symmetries in the statistical and combinatorial properties.*

*Keywords: nucleotide sequence complexity, frequency dictionary, order, Larix sibirica, Siberian larch, symmetry, transcriptome, triplet.*

*DOI: 10.17516/1997-1389-2015-8-3-278-286.*


* Corresponding author E-mail address: msad@icm.krasn.ru

# Симметрия транскриптома сибирской лиственницы

**М.Г. Садовский[а], В.В. Бирюков[б], Ю.А. Путинцева[б], Н.В. Орешкова[б,в], Е.А. Ваганов[б], К.В. Крутовский[б,г,д,е]**

*[а]Институт вычислительного моделирования СО РАН*
*Россия, 660036, Красноярск, Академгородок, 50/44*
*[б]Сибирский федеральный университет*
*Научно-образовательный центр геномных исследований*
*Россия, 660036, Красноярск, Академгородок, 50а/2*
*[в]Институт леса им. В.Н. Сукачева СО РАН*
*Россия, 660036, Красноярск, Академгородок, 50/28*
*[г]Гёттингенский университет им. Георга-Августа*
*Германия, D-37077, Геттинген, ул. Бюсгенвег, 2*
*[д]Институт общей генетики им. Н.И. Вавилова СО РАН*
*Россия, 119991, Москва, ул. Губкина, 3*
*[е]Техасский агро-инженерный университет*
*США, HFSB 305, 2138 TAMU, штат Техас 77843, г. Колледж Стейшн*

*Проанализированы структуры, выделяемые в транскриптоме лиственницы. Показано, что данный набор последовательностей обладает необычной симметрией своих статистических и комбинаторных свойств.*

*Ключевые слова: нуклеотидная последовательность, порядок, сибирская лиственница, симметрия, сложность, транскриптом, триплет, частотный словарь.*

**Introduction**

A search for an order and structuredness in bulk objects (considerably homogeneous) is one of the key scientific problems. Modern genomics (as well as linguistics) is the second to none in that respect. Indeed, amount of raw nucleotide sequence data grows daily for billions of megabytes. Those sequences are symbol sequences based mainly on the four-letter alphabet $\aleph = \{A, C, G, T\}$. For our further analysis we also assumed that neither other symbols, nor blank spaces are supposed to be found in a sequence; a sequence under consideration is also supposed to be coherent (i. e. consisting of a single piece).

We studied an order and structuredness over a set of sequences from finite alphabet $\aleph$; all these sequences represented the transcriptome of Siberian larch *(Larix sibirica* Ledeb.). Transcriptome represents sequences of expressed genes and corresponds to the mRNAs molecules isolated from biological cells or tissues.

Usually, a distance between sequences is used to find out regularities or similarities among them. Distances between sequences are most often based on sequence alignments (see Needleman and Wunsch (1970) for the first computer algorithm for aligning two sequences and Tsiligaridis (2015) for a recent review). However, there could be serious problems and constraints

in generating reasonable alignments for highly divergent sequences. Meanwhile, there are other concepts and methods that could be much more powerful than those that are based on alignments (e.g., Znamenskij, 2014, 2015), although they still need further investigation of their applicability for addressing biological problems.

Key idea in our search for a structure and order in a set of symbol sequences (transcriptome nucleotide sequences) is to translate sequences into their frequency dictionary (Bugaenko et al., 1996, 1997, 1998; Hu and Wang, 2001). There could be a number of various definitions of a frequency dictionary, but we will use the basic one that is a list of all the strings of a given length accompanied with a frequency of each string (a detailed description is given below). It is crucial that the transformation of a symbol sequence into a frequency dictionary allows us to map a set of sequences into a metric space. The latter provided us with powerful and extended tools for analysis.

We will briefly outline the concept of our study and then demonstrate the main results obtained. First, we changed each symbol sequence (that is a nucleotide sequence in the Siberian larch transcriptome set) into a frequency dictionary. Then, we studied distribution of those dictionaries in a multidimensional space trying to infer any regularities and clusters. Second, for each clustering we checked for stability of clustering. This clustering was carried out using the *K*-means technique. Third, we compared the statistical properties of the clusters identified by *K*-means and found that these clusters demonstrated a very strong symmetry in terms of the statistical properties. In brief, the clusters showed extremely low level of discrepancy in the Chargaff's second parity rule. This low discrepancy is the most intriguing fact concerning the properties of the studied transcriptome sequence set.

## Materials and Methods

### *Transcriptome nucleotide sequence data*

The transcriptome of Siberian larch was originally sequenced in the project on the whole genome sequencing of Siberian larch (Krutovsky et al., 2014). The sequence data of *L. sibirica* were obtained using Illumina MiSeq sequencer at the Laboratory of forest genomics of the Siberain Federal University. The RNA was isolated from buds (Oreshkova et al. 2015). Total number of sequences in the transcriptome set was 25 748. The shortest sequence had 128 nucleotide base pairs or bp (symbols), while the longest one had 8 512 bp. An average length of the sequences in the transcriptome was 656 bp, with the standard deviation of 565 bp. The histograms of the distribution of the transcriptome sequences entries over their length are presented in Fig. 1. Evidently, the distribution resembles quite strongly Poisson distribution. We excluded from the further consideration all the sequences entries shorter than 2000 bp to avoid a degeneracy of frequency dictionaries developed from shorter sequences. Surely, this part of the transcriptome requires special studies.

### *Frequency Dictionary*

Previously (Bugaenko et al., 1996, 1997, 1998; Hu and Wang, 2001), a frequency dictionary was proposed to be a fundamental structure of a symbol sequence. Here we introduce the definition of **frequency dictionary.** First, consider a symbol sequence $\Im$ of the length $N$ from the four-letter alphabet. Hereafter we assume that no other symbols but those from the alphabet $\aleph = \{A, C, G, T\}$ are found in a text $\Im$, and there are also no gaps in the text. The following are a few definitions:

**Definition 1:** *The* word $\omega_q = \nu_1\nu_2\nu_3\ldots\nu_{q-1}\nu_q$ *of the length q is a string occurred in the text* $\Im$.

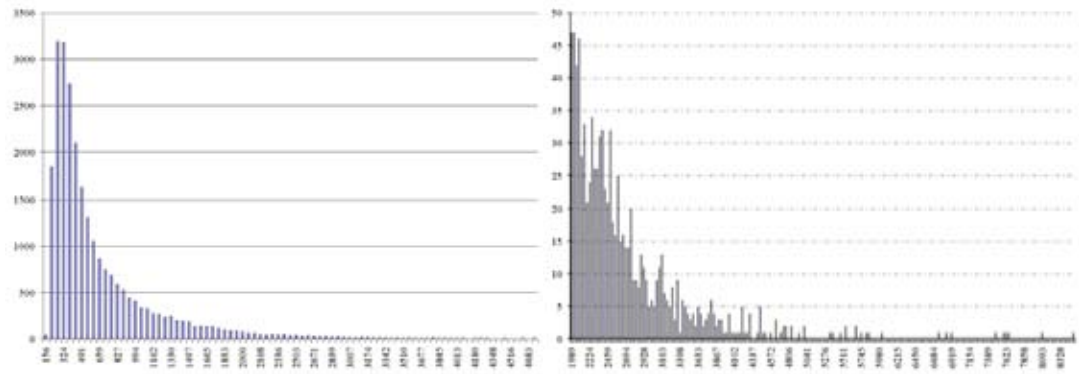

Fig. 1. Distribution of the *Larix sibirica* transcriptome sequence lengths for all sequences (left) and for sequences ≥ 2000 bp (right)

*Here $v_j$ is a symbol occupying the j-th position at the word;* $v_j \in \aleph$.

Hereafter we will consider only the words of the length 3 bp and call them *triplets.*

**Definition 2.** Frequency dictionary *W*(*q*) *is the set of all the words of the length q counted within the text* $\Im$ *so that each word is accompanied with its frequency.*

A frequency of a word ω is defined traditionally: that is the number $n_\omega$ of copies of the word divided the total number of all copies of all the words (Bugaenko et al., 1996, 1997, 1998; Hu and Wang, 2001). Frequency dictionary (namely, a dictionary *W*(3)) unambiguously maps a text $\Im$ into a 64-dimensional space, where the triplets are coordinate axes in those space, and the frequencies are the coordinates. Hence, frequency dictionary represents a short range (or meso-scale, at most) structuredness in a symbol sequence. That is the basic issue for further analysis of statistical properties of a symbol sequence representing the *L. sibirica* transcriptome. The key idea of the study was to check whether the sequences corresponding to various mRNA molecules differ in their statistical properties or not. Only the properties expressed in terms of frequency dictionaries *W*(3) will be considered.

*Clustering techniques*

We used *K*-means technique to analyze the transcriptome. *K*-means technique is a good method for analyzing data of various nature (see Fukunaga, 1990). However, the following definitions should be additionally presented beforehand:

**Definition 3:** Center *of a class developed due to K-means is the arithmetic mean of the numbers describing the objects, to be calculated within a class.*

Of course, both a *K*-means implementation and the definition of a center depend on the metrics used to do it. We used Euclidian distance as a metric in our study.

**Definition 4:** Radius *of a class is the arithmetic mean of the distances calculated from each point of a class against its center.*

High popularity of *K*-means yet does not make it free from a few following problems: (1) stability of a final distribution; (2) the number of classes determination, and (3) separability of classes. We discuss them here in more detail. Since implementation of a clustering by *K*-means starts from a random dispersion of the original data set for *K* classes, then one may not be sure that the final composition of the classes would remain the same in a new run of clustering.

Of course, one might face the situation when the final distribution is identical for any initial separation. This is the situation of the highest stability of clustering. Looking ahead, we have found a strong stability of clustering done over the transcriptome.

On the contrary, there might be a situation where any new run of the procedure brings absolutely another final distribution. This is an instability case, and such instability could be a signature of a total lack of any intrinsic structuredness in the data set. In reality, the situation could be somewhere in between. As a rule, a set of data is divided into two subsets: the former tends to yield rather stable distribution in a series of *K*-means runs, and the latter gathers the objects that always change their class attribution. These are so called volatile objects. There is no simple or evident way to deal with them. An elimination of them from the original data set may cause a loss of the stability of a new classification for the rest of a data set. Here one may assume a failure of the method; otherwise, they should be considered as a separate set to be specially studied.

An implementation of a classification through *K*-means described above is not a final step yet. One must check whether the obtained classes are distant enough. This problem is tightly related to the problem of a number of classes to be developed by *K*-means. Indeed, there is no *a priori* way to figure out the exact number of classes for *K*-means classification; it is a matter of expertise of a researcher, as a rule. Thus, an advanced technique protocol implies that one starts from a sufficiently large number of classes, and then the number is decreasing through the amalgamation of indistinguishable classes.

There are various criteria to distinguish classes. The strictest one assumes that two classes are distinguishable, if a distance between the classes is not less than the sum of those two radii. On the contrary, the weakest one requires that the greater radius must not exceed the distance between the classes. If two classes are indistinguishable, then they must be amalgamated, and *K*-means must be run again, with a new number of classes decreased by one. Thus, the advanced version of the method does not increase the number of classes and stops at the maximal set of distinguishable classes. Here we will demonstrate that the classes observed over the transcriptome in this study were highly distinguishable.

## Results

### *Classification with K-means*

We developed consequently four classifications using *K*-means technique. The number of classes varied from two to five. For each number of classes 350 runs of *K*-means were executed to study the stability of classification. The stable subsets of sequences comprising the transcriptome were determined for each classification. We assumed a classification to be stable, if not less than 95% of all runs yielded the same distribution of sequences. Moreover, the radii of classes were calculated for each classification, as well as the distances between them (see Definitions 3 and 4 above). All four classifications were very stable for 350 runs.

### *Chargaff's parity rule for two classes*

Considering the structure of two-class clustering developed over *K*-means in more detail we have to note that it was very stable (see paragraph above); besides, it yields a good separability of classes. It is a well-known fact that any symbol sequence forming a four-letter alphabet ℵ follows Chargaff's (generalized) second parity rule. The rule states that frequencies of any two words composing a complimentary palindrome should be very close. Complimentary

palindrome is a couple of words (strings) that read equally in opposite directions. In respect to the Chargaff's substitution rule it means that A ≈ T, C ≈ G (ATC ↔ GAT is an example of such palindrome). In any frequency dictionary of a thickness $q$ there always exists $\left(\frac{1}{2}\right)4^q$ couples of complimentary palindrome; in particular, for triplet frequency dictionaries there always exist 32 couples of such triplets.

Thus, we have checked the pattern of the second Chargaff's parity rule feasibility for three cases:

- within the first class identified through *K*-means;
- within the second class identified through *K*-means, and
- between these two classes.

The most surprising thing was that the second (generalized) Chargaff's parity had significantly less discrepancy in the third case (the comparison of the centers of two classes of the sequences comprising the transcriptome). Indeed, the discrepancy μ defined according to (Grebnev and Sadovsky, 2014) as

$$\mu = \frac{1}{|\Omega|}\sqrt{\sum_{\omega\in\Omega}\left(f_\omega - f_{\overline{\omega}}\right)^2} \qquad (1)$$

showed the drastically different distributions for those three cases mentioned above. Here Ω means a set of complementary palindromes (consisting of 32 entries for the cases one and two, and of 64 entries for the third case), while ω and $\overline{\omega}$ are the words making the complementary palindrome; $|X|$ means the capacity of a set $X$. The figure μ defined according to (1) looks like a distance, meanwhile it is not. It does not provide a measure between two points, but presents a discrepancy of a given dictionary. Figures 2 and 3 illustrate this fact.

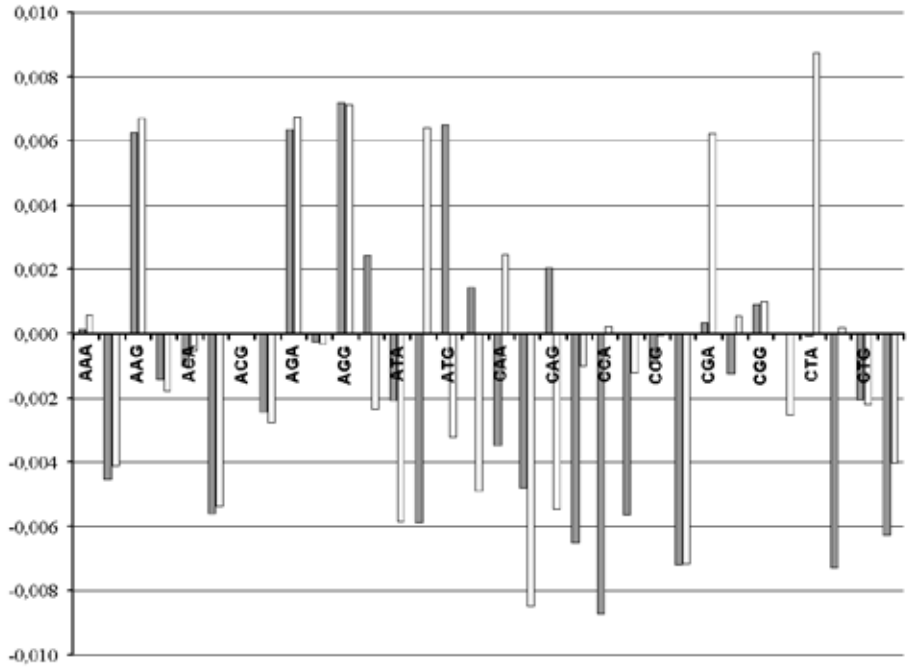


Fig. 2. The discrepancy in distribution of two classes obtained through *K*-means based on formula (1). The figure shows the differences between the frequencies of the triplets comprising the complementary palindromes: gray bars show class 1, white bars show class 2

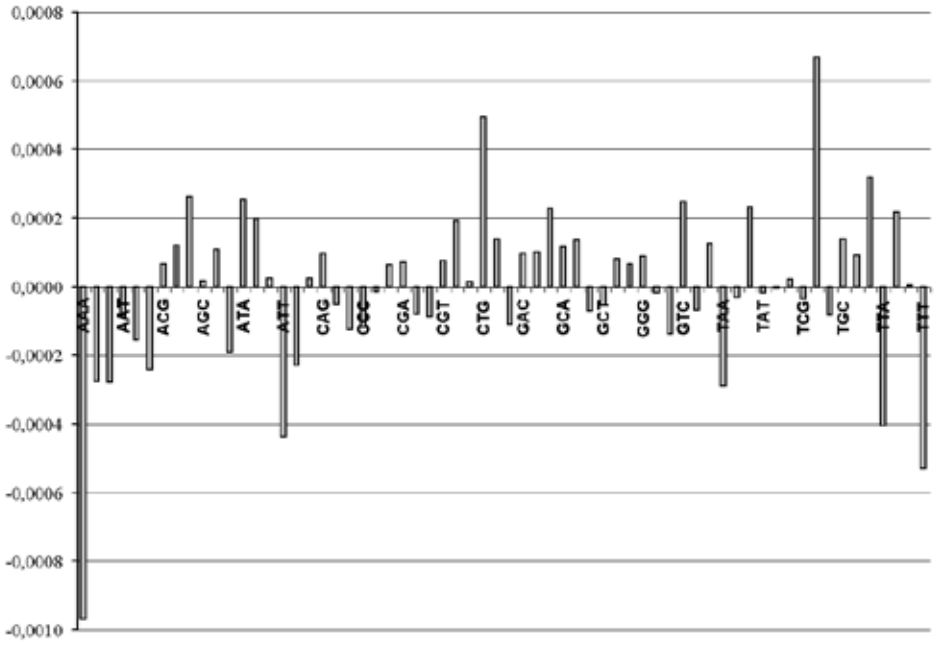


Fig. 3. The discrepancy between two classes obtained through *K*-means based on formula (1). For each couple of triplets the difference of the relevant frequencies is shown; not the significantly less values of that latter, in comparison to those shown in Fig. 2.

**Discussion**

We found out that the key issue of statistical structure of the *L. sibirica* transcriptome was a high symmetry (resembling two subsets) of the sequences comprising the transcriptome set. Obviously, the symmetry could result from the only reason: there were a lot of the pairs of "sense-antisense" type among the sequences in the set. The question is what is a source of those pairs of sequences.

Genomes of organisms of various taxonomy rank differ significantly in the discrepancy of Chargaff's generalized second parity rule based on formula (1). Mitochondria takes the leading position in terms of a level of violation of the second generalized parity rule. However, in general, there is less discrepancy of the rule for genome at a higher taxonomy rank (Grebnev and Sadovsky, 2014). Chloroplasts are, in general, the second to mitochondria.

It is unlikely that the complementary pairs would result somehow in the process of cDNA synthesis due to some technical reasons. Alternatively, such composition of the set may arise from expression of genes in opposite directions in the complementary strands of the original DNA. This hypothesis would require genes in the *L. sibirica* genome to have opposite orientation in both genomic DNA strands. Such genome arrangement is known for many (if not most) organisms, but it is mostly asymmetrical (e.g. Niu et al., 2003; Qu et al., 2010). However, the key question here is the ratio of the genes in opposite directions in the complementary strands in *L. sibirica*. A general estimation of the ratio of the opposite genes varies from 10 ÷ 1 to 1 ÷ 1, for various organisms and different authors.

It should be also stressed out that the very low discrepancy based on formula (1) observed for the centers of two classes mentioned above requires that a number of opposite genes located

in complementary strands must also overlap. Such condition is unlikely, but it does not mean that such overlapping is not possible. Evidently, the first step to verify it may consist in checking the sequences belonging to various classes (obtained due to *K*-means technique) by BLAST. This can yield a list of sequences homological to the main strand, and to the complementary one, correspondingly. Such verification may also bring another advantage: it is a common fact that BLAST is a very time-consuming procedure. A combination of *K*-means technique to figure out the tentative specific strand strings with BLAST may seriously decrease the resource demand due to the specific pre-treatment of a set to be checked. However, these issues are beyond the scope of this paper and require additional studies.

**Conclusion**

The unusual symmetry manifesting in statistical properties of triplet frequency dictionaries was found in the *L. sibirica* transcriptome. Namely, the nucleotide sequences that were transformed into the frequency dictionaries demonstrated unusually high level of the coincidence of the frequencies of (rather long) oligonucleotides with their counterparts in the opposite strand of genomic DNA. Such high coincidence may confirm an occurrence of a rather significant number of genes occupying the opposite strands and possibly also overlapping each other. It is not observed for the protein coding genes in eukaryotes, but could hypothetically occur for other gene types.

The uncertainty level may be decreased through the comprehensive analysis of all the sequences comprising the transcriptome by BLAST so that the strings belonging to opposite strands would be identified. Simultaneously, if this separation takes place, the pre-treatment for BLAST analysis using *K*-means could significantly improve the BLAST analysis due to partitioning of the strings for such analysis.

**Acknowledgements**

This study was supported by a research grant № 14.Y26.31.0004 from the Government of the Russian Federation.

## References

1. Bugaenko N. N., Gorban A. N., Sadovsky M. G. (1996) Towards the definition of information content of nucleotide sequences. Molecular Biology 30(5): 529-541.
2. Bugaenko N. N., Gorban A. N., Sadovsky M. G. (1997) The information capacity of nucleotide sequences and their fragments. Biophysics 5: 1063-1069.
3. Bugaenko N. N., Gorban A. N., Sadovsky M. G. (1998) Maximum entropy method in analysis of genetic text and measurement of its information content. Open Systems & Information Dyn. 5(2): 265-278.
4. Fukunaga K. (1990) Introduction to statistical pattern recognition. San Diego, London: Academic Press, 578 p.
5. Grebnev Ya. V., Sadovsky M. G. (2014) Chargaff's second rule and symmetry in genomes. Fundamental studies 12(5): 965-958.
6. Hu R., Wang B. (2001) Statistically significant strings are related to regulatory elements in the promoter regions of *Saccharomyces cerevisiae.* Physica A 290: 464-474.
7. Krutovsky K. V., Oreshkova N. V., Putintseva Yu. A., Ibe A. A., Deich K. O., Shilkina E. A. (2014) Preliminary results of *de novo* whole genome sequencing of Siberian larch (*Larix sibirica* Ledeb.) and Siberian stone pine (*Pinus sibirica* Du Tour.). Siberian Journal of Forest Science 1(4): 79-83.

8. Needleman S. B., Wunsch C. D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology 48(3): 443–53.
9. Niu D. K , Lin K., Zhang D.Y. (2003) Strand compositional asymmetries of nuclear DNA in eukaryotes. J. Mol. Evol. 57(3): 325-334.
10. Oreshkova N. V., Putintseva Yu. A., Kuzmin D. A., Sharov V. V., Biryukov V. V., Makolov S. V., Deych K. O., Ibe A. A., Shilkina E. A., Krutovsky K. V. (2015) Genome sequencing and assembly of Siberian larch (*Larix sibirica* Ledeb.) and Siberian pine (*Pinus sibirica* Du Tour) and preliminary transcriptome data. In: Proceedings of the 4th International Conference on Conservation of Forest Genetic Resources in Siberia. August 24-29, 2015, Barnaul, Russia, p. 127-128.
11. Qu H., Wu H., Zhang T., Zhang Z., Hu S., Yu J. (2010) Nucleotide compositional asymmetry between the leading and lagging strands of eubacterial genomes. Res. Microbiol. 161: 838-846.
12. Tsiligaridis J. (2015) Multiple sequence alignment and clustering with dot matrices, entropy, and genetic algorithms. In: Li K.-C., Jiang H., Yang L. T., Cuzzocreapp A. (eds.) Big Data: Algorithms, Analytics, and Applications. CRC Press, p. 71-88.
13. Znamenskij S. V. (2014) Modeling of the optimal sequence alignment problem. Program systems: Theory and applications 4(22): 257–267 (In Russian).
14. Znamenskij S. V. (2015) A model and algorithm for sequence alignment. Program systems: Theory and applications 1(24): 189–197.