Chapter

Revealing the Symmetry of Conifer Transcriptomes through Triplet Statistics

Sadovsky Michael, Putintseva Yulia, Biryukov Vladislav and Senashova Maria

Abstract

The novel powerful technique is used for a study of combinatorial and statistical properties of transcriptome sequences. The main approach stands on the study of distribution of nucleotide triplet frequency dictionaries obtained from the conversion of transcriptome sequences. The distribution is revealed through PCA presentation and elastic map technique. The transcriptomic data of Siberian larch (*Larix sibirica* Ledeb.) and Siberian pine (*Pinus sibirica* Du Tour) were studied. The transcriptomes exhibit unusual symmetries. The octahedral structure exhibiting rotational symmetry in transcriptome contig distribution was found for *L. sibirica*, while mirror symmetry was found for *P. sibirica*. The octahedron structure seems to be universal for plants.

Keywords: Chargaff's parity, order, structuredness, mirror symmetry, rotational symmetry

1. Introduction

A discovery of an order and new structures in genetic entities is an up-to-date scientific problem. Indeed, the amount of primary genomic data shows the daily growth for billions of megabases. The symbol sequences from four-letter alphabet $= \{A, C, G, T\}$ (with few variations in some nucleotide sequences; say, U substitutes T in RNAs).

We studied an order and structuredness over a set of sequences representing the transcriptome of Siberian larch (*Larix sibirica* Ledeb.) and Siberian pine (*Pinus sibirica* Du Tour), also known as Siberian cedar. Transcriptome represents sequences of expressed genes and corresponds to the mRNA molecule isolated from biological cells or tissues. Obviously, whether a transcriptome exhibits structuredness or not heavily depends on the concept of a structuredness to be revealed and analyzed. One may face a huge number of patterns claimed to be structural units; a number of papers report on newly discovered structures in genomes [1].

There are two approaches to discuss structuredness in a set of symbol sequences (transcriptome nucleotide sequences, in our case). The first implies that one seeks for inhomogeneities in the mutual distribution of the sequences form the ensemble under consideration. Of course, to do it, one must introduce a metrics to measure

the difference between any two sequences; there are various ways to do it [2–4]. An alignment might be such a measure [5, 6] (see also much more prominent approach presented in [7, 8]). Alternatively, the second approach implies the search for inhomogeneities within a sequence, e.g., through the comparison of the formally identified fragments of a sequence.

Regardless the specific approach to seek for structuredness, one must introduce a way to measure the difference between the objects to be analyzed. Alignment [9–11] is the most widespread approach here. An alternative idea to search a structure and order in symbol sequences is to transform them into frequency dictionary [12–15]. A frequency dictionary could be defined in various ways, but basically it is a list of all the strings of a given length accompanied with a frequency of each string (a detailed description is given below). A transformation of a symbol sequence into a frequency dictionary provides a mapping of a set of sequences into a metric space. Hence, one may apply all the tools for analysis.

As soon, as a structure in ensemble of sequences, or over a sequence is defined, the question arises toward the properties of those structures. Probably, symmetry of such structures is the most fundamental and basic one. Again, there could be various notions of the symmetry. The first concept of the symmetry aims to figure out structures that seem to remain similar, when some simple transformations in a proper space are provided. First of all, a rotational symmetry of a cluster structure [3, 4] or mirror symmetry [16, 17] must be mentioned here.

Few words should be said toward the symmetry. Here we shall consider two notions of that issue. The first is a well-known rotational, mirror, or similar symmetry observed in the distribution of the contigs converted into triplet frequency dictionary as they are distributed in the relevant Euclidean space (where the triplets are the coordinates). The second issue is measured through the proximity (or deviation) to Chargaff's parity rules, to be observed for various entities, both natural (these are contigs) and artificial (kernels or arithmetic means of the frequency of identical triplets counted over an ensemble of contigs).

2. Material and methods

2.1 Transcriptome nucleotide sequence data

The transcriptomes of Siberian larch and Siberian pine were originally sequenced under the project on the whole genome sequencing of Siberian larch [18, 19]. The sequence data of *L. sibirica* and *P. sibirica* were obtained using Illumina MiSeq sequencer at the Laboratory of Forest Genomics of the Siberian Federal University. The RNA was isolated from buds [19].

2.1.1 L. sibirica bud transcriptome

For the purposes of our study, we have selected the bud transcriptome of *L. sibirica*; we have taken into consideration the transcripts longer than 600 bp. The longest one in the transcriptome is as long as 10,795 bp, with average length $\langle L \rangle =$ 1243.4 bp and standard deviation $\sigma_{\langle L \rangle} = 717.9$ bp.

The total number of sequences in the transcriptome is 12,353 transcripts. The histograms of the distribution of the transcriptome sequence entries over their length are presented in **Figure 1**. Evidently, the distribution resembles Poisson distribution quite strongly. There are 7573 transcripts in the transcriptome bearing a single CDS (maybe in various directions). Four thousand thirty-eight transcripts

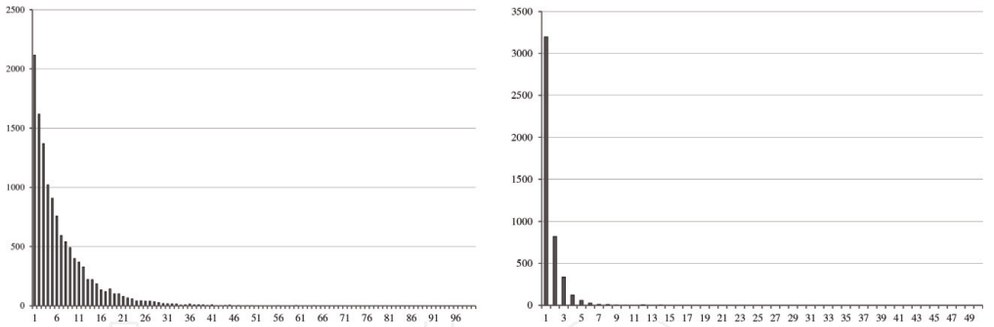

Figure 1.
Distribution of L. sibirica contigs over the length (left) and P. sibirica (right).

#	2	3	4	5	6	7	8	20
L. sibirica	3049	738	175	61	8	2	2	1
P. sibirica	962	226	41	14	3	—	—	—

#—number of CDS in a transcript.

Table 1.
Distribution of number of CDS per transcript.

have two or more CDS in them; the distribution of number of CDS in transcripts is shown in **Table 1**. Finally, in 742 transcripts no CDS have been found.

2.1.2 P. sibirica bud transcriptome

We used bud transcriptome from *Pinus sibirica* obtained from witch's broom (i.e., morphologically different part of a tree). It might be considered as a disease. Again, we have selected the transcripts longer that 600 bp that yields 4675 entries in the transcriptome, 3003 among them have a single CDS.

There are as many as 426 transcripts with no CDS detected in them. Surprisingly, there are no transcripts in the transcriptome with CDS belonging to both strands, simultaneously. The distribution of number of CDS found in a transcript is shown in **Table 1**. On the contrary to *L. sibirica* transcriptome, *P. sibirica* transcriptome contains no transcript without CDS

2.2 Triplet frequency dictionary

Triplet frequency dictionary $W_{(3,t)}$ is the list of all 64 triplets found within a sequence under consideration, where each entry (triplet) ω is assigned with the frequency f_ω of the triplet ω. The reading frame move t could be chosen arbitrary and depends on the specific problem to be solved. Everywhere further we use $t = 1$ or $t = 3$; for $t = 1$ we use the notation of W_3, unless it makes a confusion.

A frequency dictionary $W_{(3,t)}$ unambiguously maps a sequence into a point in 64-dimensional metric space. Strongly speaking, $W_{(3,t)}$ with $t>1$ maps a subsequence into the point of the metric space, not the sequence entirely; further we shall discuss this point in more detail. Next, the dimension of the space is 63, not 64; this fact follows from the linear constraint:

$$\sum_{\omega=\mathrm{AAA}}^{\mathrm{TTT}} f_\omega = 1. \tag{1}$$

This constraint allows to exclude any triplet from the analysis, thus changing 64-dimensional space for 63-dimensional, where all variables are linearly independent [20].

Formally speaking, any triplet could be excluded. Practically, one must eliminate the triplet with the least standard deviation figure determined over the set of frequencies under consideration. Indeed, suppose a triplet ω^* yields the standard deviation equal to zero, as determined over a set of dictionaries, it means, all dictionaries in the set have the same frequency, for this triplet: $f^j_{\omega^*} = const$, $\forall j$ (here j enlists the dictionaries in the set). Such invariance makes the dictionaries (and the sequences standing behind) indistinguishable, from the point of view of the triplet. The choice of a triplet with minimal standard deviation for the exclusion provides the elimination of the variable contributing least of all in distinguishability of the entities.

2.2.1 Metric choice

The list of triplets accompanied with the frequency of each entry makes frequency dictionary $W_{(3,\,t)}$; let $t = 1$, at the moment. Hence, a dictionary is a point in metric space; obviously, one may define metrics in a number of ways, in such space. For the purposes of further analysis, we use the Euclidean metrics:

$$\rho\left(W_3^{[i]}, W_3^{[j]}\right) = \sqrt{\sum_{\omega=\mathrm{AAA}}^{\mathrm{TTT}} \left(f_\omega^{[j]} - f_\omega^{[i]}\right)^2}. \tag{2}$$

Some other metrics might be used, as well. Here i and j index two different dictionaries (sequences, respectively).

2.3 Chargaff's imparity index

To begin with, we bring to mind the well-known complementarity pattern established by E. Chargaff in 1952 [21, 22]; it consists in a strong equality of A's and T's numbers (C's and G's numbers, respectively) counted over DNA molecule. Of course, some minor violations may take place due to mutations; meanwhile the accuracy of this equality is very high. This fact is also known as the first Chargaff's parity rule.

The second Chargaff's parity rule stipulates that

$$n_\mathrm{A} \approx n_\mathrm{T} \quad \text{and} \quad n_\mathrm{C} \approx n_\mathrm{G}, \tag{3}$$

if counted within a single strand. The accuracy of (3) is rather high but varies for different taxa.

Surprisingly, similar to (3) relations are observed for oligonucleotides counted over a single stand. Let us now introduce some rigorous definitions and notions.

Definition 1. Consider a string $\omega = \nu_1\nu_2...\nu_{q-1}\nu_q$ be an oligonucleotide of the length q, where ν_j is nucleotide occupying the j-th position. *Palindrome* is the word $\omega^* = \nu_1^*\nu_2^*...\nu_{q-1}^*\nu_q^*$ read equally in the opposite direction: $\nu_j = \nu_{q-j}^*$.

Definition 2. Two strings ω and $\overline{\omega}$ make the complementary palindrome, if they are read equally in the opposite directions, with respect to Chargaff's complementarity rule:

$$\mathrm{A} \Leftrightarrow \mathrm{T} \qquad \mathrm{C} \Leftrightarrow \mathrm{G}.$$

Hence, $\forall j,\ 1 \le j \le q\ \nu_j \mapsto \nu^*_{q-j+1}$. Here are some examples of complementary palindromes:

$$\mathrm{ACT} \Leftrightarrow \mathrm{AGT}, \quad \mathrm{ACTGG} \Leftrightarrow \mathrm{CCAGT}, \quad \mathrm{ACGT} \Leftrightarrow \mathrm{ACGT}.$$

So, the generalized second Chargaff's rule stipulates equality (or proximity, to be exact) of frequencies of two strings comprising complementary palindrome [23–33]. Surely, one hardly could expect to get the absolute equality of the frequencies of any two strings comprising complementary palindrome. There is a number of reasons standing behind the violation of such absolute equality; they range from purely combinatorial [25–27, 34] and/or finite sampling effect to biological peculiarities [24, 28, 30, 33].

To reveal the difference between genetic entities or biological objects, one must introduce a measure of the violation of the generalized second Chargaff's rule; one may do it in various ways; we use the discrepancy index:

$$\mu\left(W_q^{[i]}, W_q^{[j]}\right) = 4^{-q} \cdot \sqrt{\sum_{\omega \in \Omega} \left(f_\omega - f_{\overline{\omega}}\right)^2}. \tag{4}$$

Here Ω is the set of strings of the length q observed in two sequences (i and j, respectively), ω enlists all the strings, and $\overline{\omega}$ is the string complementary palindromic to ω. Normalization factor 4^{-q} is introduced to equalize the figures (4) observed for various q.

The index (4) measures the discrepancy between two dictionaries ($W_q^{[i]}$ and $W_q^{[j]}$). Meanwhile, this index could be applied for a single frequency dictionary W_q:

$$\mu^\star\left(W_q\right) = 2 \cdot 4^{-q} \cdot \sqrt{\sum_{\omega \in \Omega^\star} \left(f_\omega - f_{\omega^\star}\right)^2}. \tag{5}$$

Here the complementary palindromic couples are combined from the strings belonging to the same frequency dictionary W_q.

The discrepancy measure (4) looks like Euclidean distance, while it is not. More exactly, it could be considered as a metrics in Euclidean space. To do it, one must reconsider a point in a couple, changing it for the dual one that is a complementary palindrome.

The inner discrepancy measure (5) definitely is not a distance, since it characterizes a single object, not a couple.

2.4 $W_{(3,\,3)}$ and W_3 dictionaries

This is a very common fact that a genome comprises coding and noncoding regions. Basically, they differ in the statistical properties manifested in triplet frequency dictionaries. One might detect some minor difference in W_3 composition developed for coding vs. noncoding regions. Significantly greater difference between these two types of genome parts is observed for $W_{(3,\,3)}$ dictionaries [2–4].

Dictionary W_3 is uniformly defined, for any sequence. The situation differs for $W_{(3,\,3)}$ dictionaries. Consider a sequence $\mathfrak{L}$ of the length N. Starting to cover the sequence with the frames of the length 3 moving along the sequence with the step 3, one may get three different dictionaries, in dependence to the location of the start point. The starts may be located at the first nucleotide of a sequence, at the second nucleotide, and at the third nucleotide; thus, three different triplet frequency dictionaries $W_{(3,\,3)}$ could be obtained.

The key difference between coding and noncoding regions consists in the deviations between these three dictionaries. In other words, let the sequence $\mathfrak{L}$ falls entirely into a noncoding region of a genome. One may develop three triplet frequency dictionaries $W^{[j]}_{(3,\,3)}$, $0 \le j \le 2$ corresponding to three positions of the reading frame shift (these are 0, 1, and 2). The key issue is that these three dictionaries:

1. Differ significantly if developed for coding and noncoding regions.

2. Differ each other, if developed for a coding region.

3. Differ between them negligibly, if developed for a noncoding region.

In other words, consider a set $\hat{W}^{[j]}_{(3,\,3)}$, $0 \le j \le 2$ developed over a noncoding region and a set $\tilde{W}^{[j]}_{(3,\,3)}$, $0 \le j \le 2$ developed over a coding region. Then, $\forall j$ the difference between $\hat{W}^{[j]}_{(3,\,3)}$ is rather small, when expressed in any way (as Euclidean distance, entropy, mutual entropy, etc.; see also [7, 8]), but the difference between $\tilde{W}^{[j]}_{(3,\,3)}$ is significantly greater. Besides, $\forall i, j$ the difference between $\tilde{W}^{[i]}_{(3,\,3)}$ and $\hat{W}^{[j]}_{(3,\,3)}$ manifests apparently. These deviations in statistical properties of such triplet frequency stand behind the *Hidden Markov Model* methodology [35, 36].

We shall explore structuredness in transcriptomes through the analysis of those triplet dictionaries developed over the individual transcripts.

2.5 Relative phase

To reveal the inner structuredness of a (bacterial) genome, Gorban and coauthors have introduced special construction that might be called *tiling* [2–4]. The idea was to cover a genome (considered as a symbol sequence from) with a set of overlapping and ordered windows called tiles. All tiles are of the same length L ($L = 603$ in [2–4, 16, 17]); the tiles are located along a sequence with the permanent step P. In the papers mentioned above, $P = 11$, and the choice of the specific figures of L and P is determined by the specific task of a research.

A subsequence identified by a specific tile is then converted into frequency dictionary $W_{(3,\,3)}$, and the inner structuredness of a genome is represented through the distribution of the points corresponding to tiles, in 64-dimensional (or 63-dimensional) metric space.

This structuredness is basically determined by the so-called *relative phase* of a tile. It may:

1. Fall completely into a coding region.

2. Fall completely outside a coding region.

3. Contain a border between coding and noncoding regions.

In any chance, the relative phase indicates whether the start of a tile coincides with a start of a coding region or not. There are following combinations determining the relative phase index:

1. Start of a coding region coincides to the start of a tile. In this case relative phase $\delta = 0$.

2. Start of a coding region does not coincide to the start of a tile, and the reminder of the division of the distance (expressed in number of nucleotides) from the start of the tile, and the start of coding region is 1. Then $\delta = 1$ in this case.

3. Finally, the start of a coding region falling inside the tile does not coincide to the start of a tile, and the remainder is 2. Then $\delta = 2$ in this case.

For any tile covering a noncoding region, $\delta = 4$, by definition.

It should be stressed that genes (or coding regions) may take place in opposite strands; in such capacity, the relative phase index must be defined for leading strand and lagging one, separately, where the remainder of the division must be determined for the difference between the last symbol of a tile and the last nucleotide of a gene annotated in a sequence as located in the lagging strand. Thus, seven figures of the relative phase index δ are possible: F_0, F_1, and F_2 for the tiles containing coding regions from the leading strand; B_0, B_1, and B_2 for the tiles containing coding regions from the lagging strand; and, finally, J labeling the tiles covering noncoding regions, only.

For genome tiling (see [2–4, 16, 17]), the labeling of tiles with the relative phase index is based on genome annotation.

2.5.1 Transcriptome relative phase

The situation is slightly different for transcriptome (and the transcriptomes of *L. sibirica* Ledeb. and *P. sibirica* Du Tour, specifically). First of all, we did not develop any tiling, for transcripts; reciprocally, the transcripts themselves have been considered as tiles. It means that each transcript was converted into $W_{(3,\,3)}$ frequency dictionary as a whole, with no dissection into tiles.

Each frequency dictionary corresponding to a specific transcript was labeled with relative phase index; the labeling procedure was pretty close to that one described above, with few exceptions. We used TransDecoder™ software to find the start of a coding region within a transcript, as well as the strand location of CDS.

The relative phase index for transcripts containing a single CDS was determined in completely the same way, as described above. The transcripts bearing no CDS, if any, have been labeled with index J. Finally, the problem arose from the transcripts bearing several CDS: obviously, a relative phase index is defined ambiguously for such transcripts. In such capacity, we labeled the transcripts with multiple CDS with special figure M of the relative phase index.

Finally, we have calculated the standard deviation for each triplet, over the entire set of transcripts; that is CGT with $\sigma_{\mathrm{CGT}} = 0.005586$, so we excluded this triplet from the set of variables to cluster the transcripts. Reciprocally, the triplet with $\sigma_{\mathrm{TGA}} = 0.014924$ yields the maximal figure of the standard deviation.

Similar figures determined for *P. sibirica* are $\sigma_{\mathrm{GCG}} = 0.005658$ and $\sigma_{\mathrm{TGA}} = 0.014936$, correspondingly; the former stands for the minimal standard deviation figure, and the latter stands for the maximal one. Hence, in cedar transcriptome, we have excluded GCG triplet. Remarkably, the triplets with the largest standard deviation figures coincide, for these two genetic entities.

3. Results

Previously, seven cluster symmetric patterns have been reported [2–4], in bacterial genomes. Later, similar (but not equivalent) structures were found in chloroplast genomes [16, 17]. First of all, the tiles corresponding to specific relative phase

tend to aggregate into clusters apparently seen in the projection into three principal components with the largest eigenvalues. The points corresponding to specific strand (either leading or a lagging one) perform a triangle, in the frequency space; the points corresponding to noncoding regions tend to gather into a ball-like structure located in the central part of the pattern.

The patterns described in [2–4, 16, 17] are provided by the interplay of two triangles and the central ball. The triangles comprise the points corresponding to specific strand. There are two basic symmetries found in these triangles: the former is a shift (rotational) symmetry peculiar for bacterial genomes [2–4], and the latter is mirror symmetry peculiar for chloroplasts [16, 17]. The ball comprise the points corresponding to the tiles with noncoding regions inside (chloroplast genomes have one more cluster called *tail*; meanwhile, it is not important at the moment).

Whether a pattern would have four or seven clusters depends on GC content of a genome, for bacteria [2–4]. This figure almost completely determines the mutual location of the planes comprising the triangles formed by the clusters belonging to the same strand. There are some exclusions from this rule, for cyanobacteria. Chloroplasts exhibit mirror symmetry in the strand-specific triangles, so they always have a four-beam structure, where the triangles occupy the same plane with obligatory coincidence of F_2 and B_2 phases [16, 17].

3.1 Phase index coloring agreement

To make the presentation of results clearer, let us fix the color and label mark usage for transcripts to be shown in figures everywhere further. Indeed, we should distinguish eight different phases in the figures: F_0, B_0, F_1, B_1, F_2, B_2, *mult*, and *noCDS*.

To do that, we shall use the following labels: all phases of F_0 through F_2 of transcripts from the leading strand are marked with triangles; all phases of B_0 through B_2 of transcripts from the lagging strand are marked with diamonds; *mult* transcripts are marked with teal squares; finally, the transcripts where no CDS have been found are labeled with brown circles.

Besides, the relative phases of single CDS transcripts are colored in the following manner: F_0 is purple triangle, F_1 is lime triangle, and F_2 is yellow triangle; reciprocally, B_0 is magenta diamond, B_1 is azure diamond, and B_2 is sand diamond.

We should say few words concerning the distribution of the transcripts with several CDS detected in them. For both transcriptomes, the distribution of such transcripts in the 63-dimensional space seems to be very homogeneous; in other words, these transcripts do not form any specific cluster, neither they are attracted to any other given one provided by the transcripts with specific (and unambiguous) relative phase index. The same is true for both studied transcriptomes. Later we discuss this point in more detail, while here we fix that the points representing such multi-CDS transcripts are erased from the pictures illustrating the results.

Thus, the clusters formed by transcripts of the same relative phase index are located in two parallel planes (in the space of three principal components with the largest eigenvalues). This observation holds true for *L. sibirica* transcriptome, while *P. sibirica* transcriptome exhibits some deviations from this pattern. We should discuss it later in more detail.

3.2 *L. sibirica* transcriptome octahedron

Unlike the tiles developed for a genome, the transcripts of a transcriptome exhibit an ultimate pattern, that is, octahedron. The rectangular triangles, ΔABC and $\Delta\alpha\beta\gamma$, in **Figure 2** occupy the position in two orthogonal planes. Note, these

triangles do not comprise the clusters from the same strand; on the contrary, phases over the octahedron are distributed in the manner shown in **Figure 2** (right).

Figure 3 shows the distribution of *L. sibirica* transcripts with relative phase values ranging from F_0 to B_2; they are colored as described above. This is the distribution in 63-dimensional space (see Section 2.5.1) shown as the projection into two-dimensional plane determined by the first and the second principal components (**Figure 3**, left) and by the second and the third principal components (**Figure 3**, right); this right image is rotated for $\pi/4$ angle clockwise.

The transcriptome shown in this figure exhibits clear and unambiguous octahedral pattern in cluster location. It is evident that F_0 to F_2 phases lay out in a plane and vice versa: the phases from the lagging strand are also laid out in a plane, and these two planes are parallel. It should be stressed that this pattern is observed in the metric space defined by the eigenvectors of the covariation matrix; in other words, the clear and apparent octahedron pattern is observed in affinely transformed space, not in the original one determined by triplet frequency.

Let us now consider the distribution of the points corresponding to *noCDS* and *mult* indexes. These two types of sequences differ drastically, in terms of their dispersion over the pattern. The transcripts bearing several CDS (see **Table 1**) are rather long. The distribution of $W_{(3,\,3)}$ of such transcripts is shown in **Figure 4**; it should be stressed that this is the mutual distribution of all the points, with the

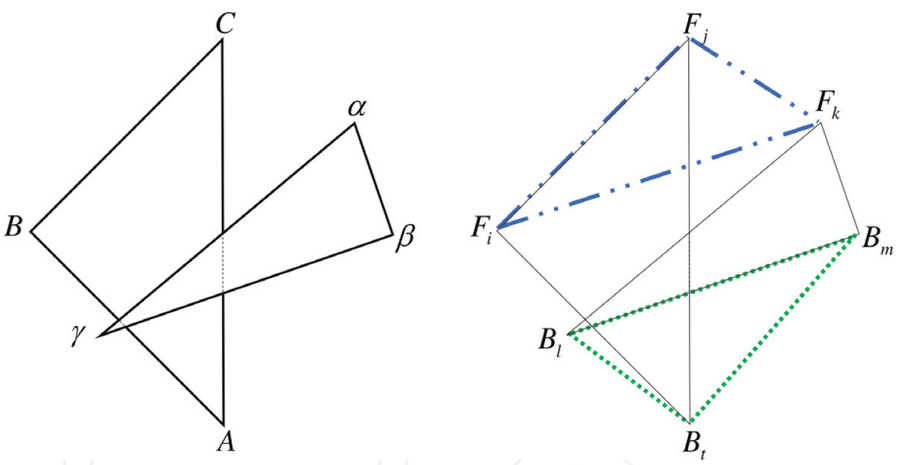

Figure 2.
Typical distribution of L. sibirica transcripts in 63-dimensional space.

Figure 3.
The distribution of L. sibirica transcripts; phases noCDS and mult are erased.

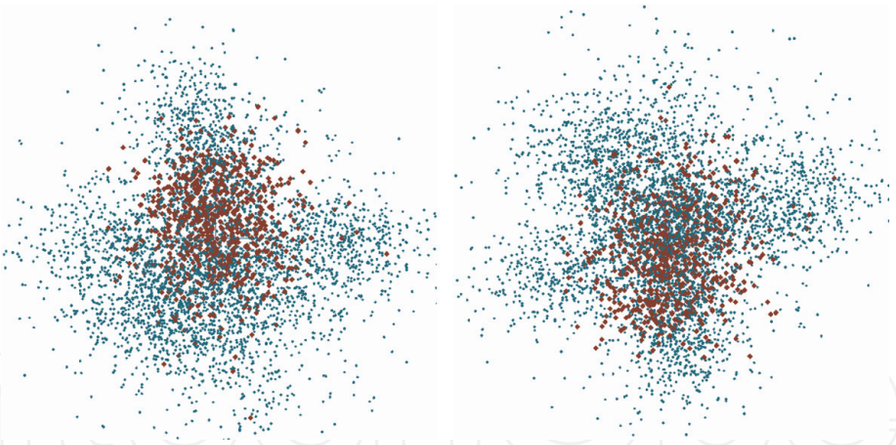

Figure 4.
The distribution of L. sibirica transcripts with noCDS (brown diamonds) and mult sequences (teal circles). The axes are directed in the same way, as in ***Figure 3****.*

complete set of phase indexes; the only point in this Figure is that the points corresponding to phases F_0 through B_2 are erased.

Also, this figure shows the distribution of the transcripts where no CDS have been found (brown circles). The cluster comprising these transcripts is rather remarkable: the transcripts where no CDS have been found behave themselves (in terms of clustering in 63-dimensional triplet frequency space) pretty close to the fragments falling completely into noncoding regions of a genome, when a complete genome is sliced into a set of tiles [2–4, 16, 17]. This observation indirectly (while rather hard) proves the total lack of any CDS in such sequences; otherwise, the corresponding frequency dictionaries never could be gathered in a ball centered at the pattern.

The transcripts with several CDS inside are distributed over the pattern almost homogeneously, including the central spot where the transcripts without CDS are concentrated. Apparently, this fact follows from the multiplicity of CDS in these transcripts: an interplay of different CDS located within a transcript may yield an effective value of its *phase* index ranging from F_0 to B_2, and the impact of those CDS is expected to be rather random.

3.3 *P. sibirica* transcriptome octahedron

Let us now focus on the peculiarities of the transcriptome of *P. sibirica*. First of all, this transcriptome (at least, the part taken into analysis) is less abundant, in comparison to *L. sibirica* transcriptome. This fact may impact the pattern of the triplet frequency dictionary distribution, while one may expect the effect to be negligible, since the length distribution of the transcripts of *P. sibirica* is similar to that one observed for *L. sibirica* (see **Figure 1**) and the portion of multi-CDS transcripts in these two transcriptomes are quite similar (see **Table 1**).

To begin with, **Figure 5** shows the clustering pattern observed for this transcriptome; the technology of the development of the pattern is absolutely the same, as in **Figures 3** and 4. The strongest difference between this transcriptome and the *L. sibirica* one consists in the significant deformation of the octahedron observed over *P. sibirica* transcriptome; **Figure 6** illustrates this point.

At the first glance, the pattern shown in **Figure 5** looks like a tetrahedron, while it is not. In proper projection, the pattern looks like a hexagon; adding the subset of multi-CDS transcripts, one gets the same pattern almost homogeneously covered by the point corresponding to the subset.

Figure 5.
The distribution of P. sibirica transcripts; the phase mult is erased.

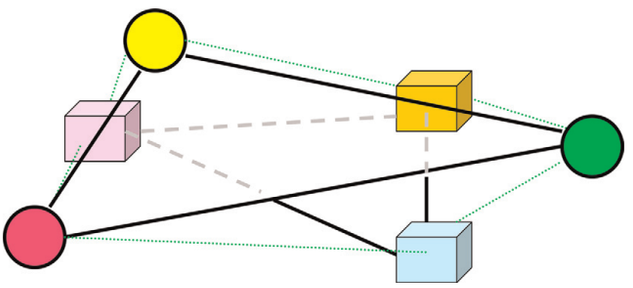

Figure 6.
The deformation of P. sibirica transcriptome. Balls are the clusters of F-strand, and boxes are the clusters of B-strand; coloring follows the layout described above (see Section 3.1).

4. Discussion

The patterns provided by the distribution of considerably short fragments of a genome may tell a lot to a researcher [2–4, 16, 17]. For bacteria, GC content seems to be the key factor determining the details of the pattern [2–4]. That is not so for chloroplasts, mitochondria, and cyanobacteria [16, 17]. The results presented above show that GC content has nothing to do with a pattern observed over a transcriptome. Hence, a question arises toward the key factor determining the specific type of a pattern. Yet, there is no simple and brief answer, while Chargaff's parity rule discrepancy may be quite informative here.

We have determined Chargaff's rule discrepancy measure (5) figure $\mu^\star$ for all six clusters observed in *L. sibirica* and *P. sibirica* transcriptomes; **Table 2** shows them. The variation of these figures $\mu^\star$ is very smooth, and the clusters are pretty close to each other, in terms of the discrepancy μ (see Eq. (5)). This fact opposes to similar observations carried out over bacterial, chloroplast, and mitochondrial genomes [16, 17]: these later exhibit significant (more than 10 times) difference in the discrepancy figures calculated for the clusters. It should be said that, unlike transcriptomes, chloroplast genomes exhibit three-beam patterns, where a beam (i.e., a cluster) comprises the fragments from forward and backward strands, simultaneously. There is no such combination, for transcriptomes.

Let us now focus on a few more details on Chargaff's imparity index, itself. The index value differs for different length q of words. Thus, a question arises toward the reference figures for this index. Suppose, the index is determined over the frequency dictionaries derived from both strands; in such capacity, it must be equal to zero.

	Relative phases					
Transcriptome	F_0	F_1	F_2	B_0	B_1	B_2
L. sibirica	0.00129	0.00169	0.00144	0.00160	0.00133	0.00123
P. sibirica	0.00122	0.00154	0.00144	0.00150	0.00135	0.00131
L. sibirica	0.12904	0.22707	0.06629	0.06774	0.09674	0.06201
P. sibirica	0.06944	0.07185	0.07023	0.07163	0.07559	0.07712

Table 2.
Discrepancy measure (5) figures $\mu^\star$ for two transcriptomes (upper part) and cluster radii, for the same phases (lower part).

Calculating the index (4) over a single strand, one may clearly understand to what extend a strand looks like the opposite one, in terms of the word frequency [23–25]. For random non-correlated sequence with $f_A = f_T$ and $f_C = f_G$ $\left(\mu_q = 0\right)$. Hence, $\forall_q . \mu_q$ figures remain the same, if the discrepancy μ_1 is fixed [23].

Unlike $\mu^\star$ figures, the radii of these six clusters exhibit quite diverse behavior. The radius of a cluster is an average distance from the center (that is arithmetic mean) determined over the cluster to each point from the cluster. Lower part of **Table 1** shows the radii figures. The radii figures are apparently different, for the transcriptomes under consideration. F_0 and F_1 phases for *L. sibirica* show extremely high values. These figures may not be explained by the excess of the cluster abundance of *L. sibirica* in comparison to *P. sibirica*. Again, the variation of the radii for *L. sibirica* is evidently greater than for *P. sibirica*, and this fact correlates to the mirror symmetry of *P. sibirica* transcriptome, since it is typical for simpler and less diverse genetic system.

Inter-cluster discrepancy measure μ is of great interest, for both cases; **Table 3** shows these indexes. Careful examination of **Table 3** allows to identify three couples of relative phase indexes with distinctively lower figure of (4), namely, the couples:

$$F_0 \Leftrightarrow B_2 \qquad F_1 \Leftrightarrow B_0 \qquad F_2 \Leftrightarrow B_1. \tag{6}$$

			L. sibirica		
Phase index	F_1	F_2	B_0	B_1	B_2
F_0	0.00095	0.00111	0.00098	0.00101	0.00007
F_1		0.00094	0.00008	0.00109	0.00102
F_2			0.00111	0.00009	0.00105
B_0				0.00090	0.00090
B_1					0.00105
			P. sibirica		
F_0	0.00091	0.00105	0.00096	0.00028	0.00099
F_1		0.00094	0.00008	0.00107	0.00105
F_2			0.00112	0.00110	0.00016
B_0				0.00098	0.00087
B_1					0.00105

Table 3.
Discrepancy measure (4) figures μ determined within each of the two transcriptomes.

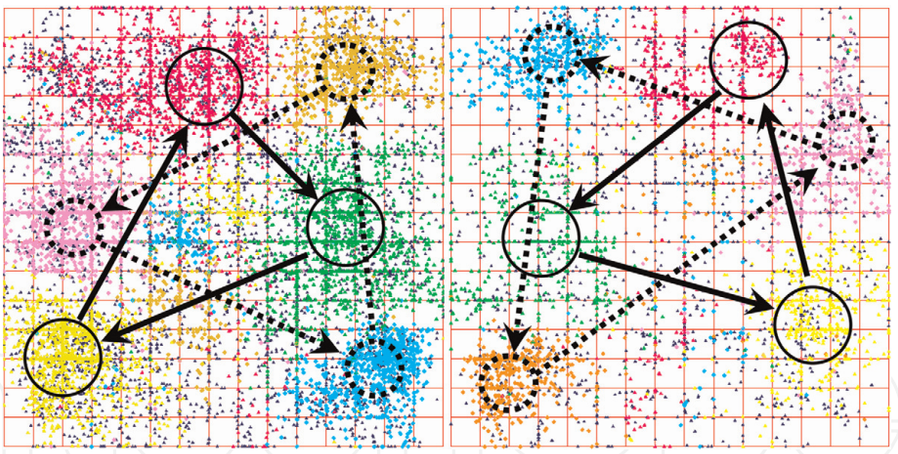

Figure 7.
Mirror (left) symmetry in L. sibirica transcriptome vs. shift symmetry (right) in P. sibirica transcriptome. Solid circles and solid arrows correspond to F phases, while dashed ones correspond to B phases.

Evidently, the phases in these couples yield two different types of symmetry: the first one is shift, and the second symmetry is mirror. The situation is opposite for *P. sibirica* transcriptome: the couples with the least Chargaff's discrepancy measure (4) are the following:

$$F_0 \Leftrightarrow B_1 \qquad F_1 \Leftrightarrow B_0 \qquad F_2 \Leftrightarrow B_2. \tag{7}$$

To make the situation with symmetries clear, we show the clusters over the elastic map shown in the so-called *inner coordinates*; **Figure 7** presents the transcriptomes.

Such mirror symmetry has been previously reported for chloroplast genomes [16, 17] (see also [23, 37, 38]); yet, there were no other but the chloroplast genomes exhibiting such mirror symmetry, and *L. sibirica* transcriptome is the next one in this point.

Definitely, the coincidence of these two symmetrical patterns does not mean that *L. sibirica* transcriptome is identical to a chloroplast genome in all other properties. Probably, plants differ from other eukaryotic organisms and bacteria in the symmetry type; currently, no eukaryotic genome is found with mirror symmetry. Shift symmetry observed for *P. sibirica* transcriptome poses a question toward the origin of the symmetry type change: whether it results from some essential biological difference between these two pine species or it is a manifestation of the genomic transformation in witch's broom cells. To answer the question, more studies are necessary.

The most amazing thing in transcriptome statistical properties is that it yields an octahedral pattern, unlike bacteria, organelle, and other genetic entities (say, yeast genomes). Another point is that the pattern does not depend on the length of transcripts taken into consideration: we have examined separately the subsets of transcripts as long as $200 \le N \le 600$ bp, $600 \le N \le 2500$ bp, and those longer 3000 bp. All these subsets yield similar pattern, with very minor variation mainly manifesting in cluster density.

One can easily see two major peculiarities differing a transcriptome from the sets of tiles described above (see [2–4, 16, 17] for details). These are:

- Total absence of the (rather extended) noncoding regions.

- Elimination of introns from the statistical analysis of sequences.

Of course, the first item from this list is quite arguable: a number of transcripts where no CDS has been detected bring a direct and unambiguous disproof of it. Thus, the question arises, whether these transcripts are similar, in some sense, to the fragments of genome comprising purely noncoding regions of the latter.

We have examined the first hypothesis through the simulation of noncoding regions. To do that, we have added a number of $W_{(3,\,3)}$ frequency dictionaries obtained from the tiles covering the noncoding parts of genomes of several other organisms. All the tiles were as long as 603 bp and contained noncoding regions, exclusively. The number of dictionaries (the points, in other words) varied from one third to one half of the total number of transcripts in the set. By assumption, this addition simulated a genome.

Upon addition, we expected to see a pattern similar to that one observed in bacteria, organelle, or other eukaryotic organisms; the octahedron pattern appeared to be stronger. **Figure 8** obviously disproves this hypothesis: it shows the same transcriptome (*L. sibirica*) with eliminated transcripts bearing no CDS, where a set of $W_{(3,\,3)}$ dictionaries borrowed from three different genomes is added,

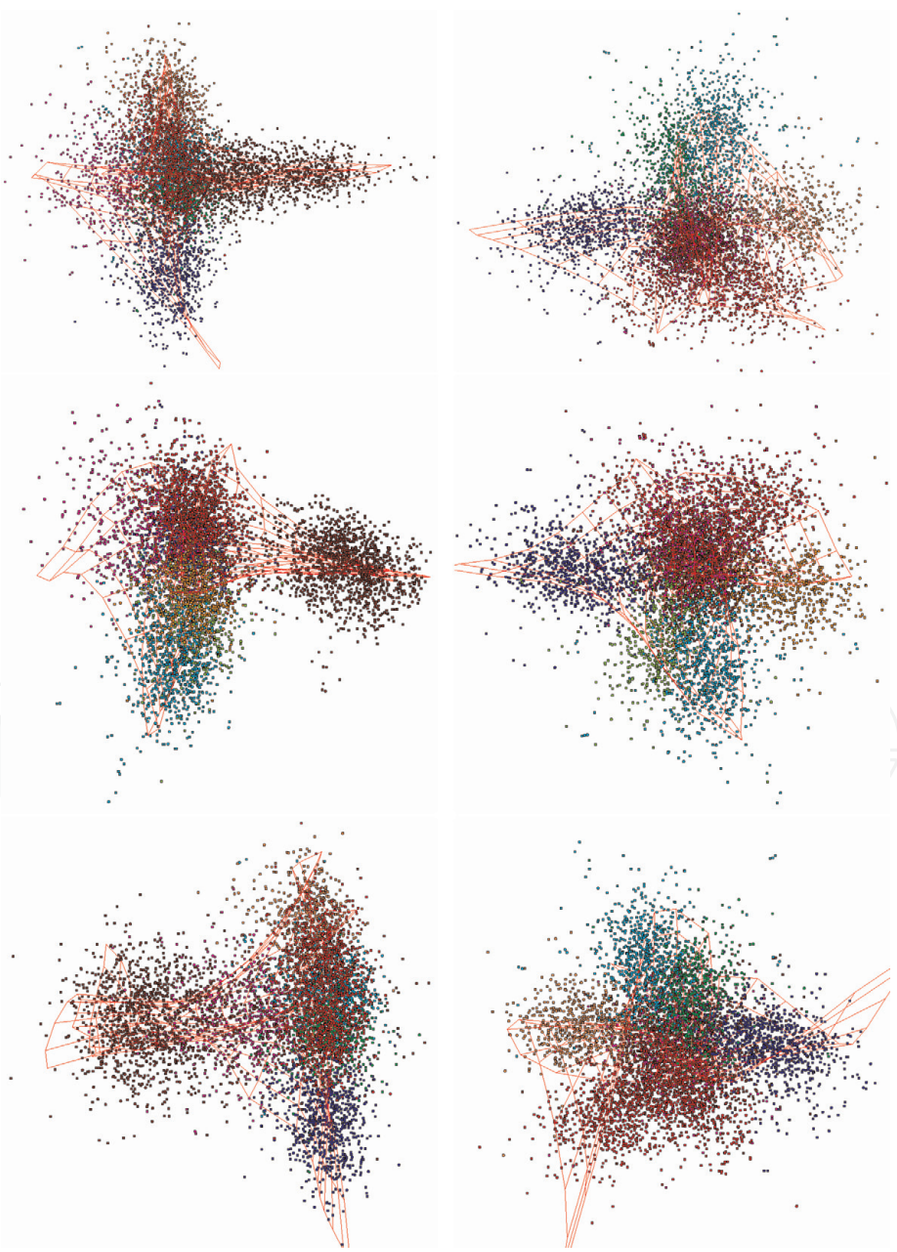

Figure 8.
Three noncoding data points added to L. sibirica transcriptome; nothing happened.

consequently. Obviously, such simulation of a genome does not break down the observed pattern of transcript distribution. Yet, one more option should be examined: what happens if the natural noncoding regions are used to simulate a genome? In other words, the pattern might be sensitive to the noncoding regions from the original genome, only, This point still awaits for examination.

The impact of introns on the alteration of the observed pattern is less evident. Moreover, one faces greater difficulties in revealing it. One might want to compare the distributions developed over $W_{(3,3)}$ and W_3 dictionaries, in this case; yet, this problem needs careful investigation and falls beyond the scope of this paper.

5. Conclusions

Systematic comparison of (rather short) fragments of permanent length formally identified within a genome reveals a symmetry in the distribution of the triplet frequency dictionaries obtained over those fragments; originally this effect has been found on bacterial genomes. Later similar (while rather different in a number of essential details) behavior has been found for chloroplasts and mitochondria genomes. The general pattern of the distribution looks like a superposition of two triangles where the vertices correspond to the fragments of the same relative phase. In simple words, it corresponds to a reading frame shift, in case of a translation-like processing of DNA sequence.

A transcriptome itself might be considered as a set of those fragments, with few exclusions. Firstly, the lengths of transcripts are different and may affect the expected pattern. Secondly, there are no fragments in a transcriptome corresponding to those obtained from noncoding (intergenic) regions of a genome. This fact results in ultimate possible configuration of the clusters corresponding to the transcripts with the same relative phase index, that is, octahedron. All these patterns could be seen in the space of three principal components with the largest eigenvalues. The *L. sibirica* transcriptome yields almost perfect octahedral pattern, while the *P. sibirica* transcriptome differs rather significantly, with planes comprising the clusters from the same strand to be located almost in parallel. This deformation might result from the biology: we studied the *P. sibirica* transcriptome obtained not from a normal tree, but from a witch's broom bud; the latter is known for extremely deviated morphology that may not avoid serious genetic alteration in its genome.

Acknowledgements

The data used in this study were obtained under the grant 14.Y26.31.0004 from the Russian Government. The authors also thank Serafima Novikova from Siberian Federal University for the helpful discussion.

Conflict of interest

The authors declare no conflict of interest.

Author details

Sadovsky Michael[1,2*], Putintseva Yulia[2], Biryukov Vladislav[2] and Senashova Maria[1]

1 Institute of Computational Modeling SB RAS, Krasnoyarsk, Russia

2 Siberian Federal University, Institute of Fundamental Biology and Biotechnology, Krasnoyarsk, Russia

*Address all correspondence to: msad@icm.krasn.ru

References

[1] Weiss LE, Naor T, Shechtman Y. Observing DNA in live cells. Biochemical Society Transactions. 2018; **46**(3):729-740

[2] Gorban AN, Popova TG, Zinovyev AY. Codon usage trajectories and 7-cluster structure of 143 complete bacterial genomic sequences. Physica A: Statistical Mechanics and its Applications. 2005;**353**:365-387

[3] Gorban AN, Popova TG, Zinovyev AY. Seven clusters in genomic triplet distributions. In Silico Biology. 2003;**3**(4):471-482

[4] Gorban AN, Popova TG, Zinovyev AY. Four basic symmetry types in the universal 7-cluster structure of microbial genomic sequences. In Silico Biology. 2005;**5**(3):265-282

[5] Chu KH, Qi J, Yu ZG, Anh V. Origin and phylogeny of chloroplasts revealed by a simple correlation analysis of complete genomes. Molecular Biology and Evolution. 2004;**1**:200-206

[6] Tsiligaridis J. Multiple sequence alignment and clustering with dot matrices, entropy, and genetic algorithms. In: Li K-C, Jiang H, Yang LT, Cuzzocreapp A, editors. Chapter 4 in Big Data: Algorithms, Analytics, and Applications. CRC Press; 2015. pp. 71-88

[7] Znamenskij SV. Modeling of the optimal sequence alignment problem. Program Systems: Theory and Applications. 2014;**4**(22):257-267 (in Russian)

[8] Znamenskij SV. A model and algorithm for sequence alignment. Program Systems: Theory and Applications. 2015;**1**(24):189-197

[9] Antipov D, Raiko M, Lapidus A, Pevzner PA. Plasmid detection and assembly in genomic and metagenomic datasets. Genome Research. 2019;**26**(9): 961-968

[10] Vignesh U, Parvathi R. Biological Big Data analysis and visualization: A survey. In: Biotechnology: Concepts, Methodologies, Tools, and Applications. IGI Global; 2019. pp. 653-665

[11] Kaur S, Kaur S, Sood SK. Proposed better sequence alignment for identification of organisms using DNA barcode. In: Innovations in Computational Intelligence. Singapore: Springer; 2018. pp. 115-150

[12] Bugaenko NN, Gorban AN, Sadovsky MG. Towards the definition of information content of nucleotide sequences. Molecular Biology. 1996; **30**(5):529-541 (in Russian)

[13] Bugaenko NN, Gorban AN, Sadovsky MG. The information capacity of nucleotide sequences and their fragments. Biophysics. 1997;**5**: 1063-1069 (in Russian)

[14] Bugaenko NN, Gorban AN, Sadovsky MG. Maximum entropy method in analysis of genetic text and measurement of its information content. Open Systems and Information Dynamics. 1998;**5**(2):265-278

[15] Hu R, Wang B. Statistically significant strings are related to regulatory elements in the promoter regions of *Saccharomyces cerevisiae*. Physica A: Statistical Mechanics and its Applications. 2001;**290**:464-474

[16] Sadovsky MG, Senashova MY, Malyshev AV. Chloroplast genomes exhibit eight-cluster structuredness and mirror symmetry. In: Rojas I, Ortuño F, editors. Bioinformatics and Biomedical Engineering. Cham: Springer International Publishing; 2018. pp. 186-196. LNBI 10813

[17] Sadovsky MG, Senashova MY, Putintseva YA. Chapter 2. Eight clusters, synchrony of evolution and unique symmetry in chloroplast genomes: The offering from triplets. In: Chloroplasts and Cytoplasm: Structure and Functions. Nova Science Publishers, Inc.; 2018. pp. 25-95

[18] Krutovsky KV, Oreshkova NV, Putintseva YA, Ibe AA, Deich KO, Shilkina EA. 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. 2014;**1**(4):79-83

[19] Oreshkova NV, Putintseva YA, Kuzmin DA, Sharov VV, Biryukov VV, Makolov SV, et al. 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: Barnaul State university; 2015. pp. 127-128

[20] Fukunaga K. Introduction to Statistical Pattern Recognition. London, Berlin, Heidelberg: Academic Press; 1990. pp. 1-625

[21] Elson D, Chargaff E. On the deoxyribonucleic acid content of sea urchin gametes. Experientia. 1952;**8**(4): 143-145

[22] Chargaff E, Lipshitz R, Green C. Composition of the deoxypentose nucleic acids of four genera of sea-urchin. The Journal of Biological Chemistry. 1952;**195**(1):155-160

[23] Grebnev YV, Sadovsky MG. Chargaff's second rule and symmetry in genomes. Fundamental Studies. 2014; **12**(5):965-968 (in Russian)

[24] Sánchez J, José MV. Analysis of bilateral inverse symmetry in whole bacterial chromosomes. Biochemical and Biophysical Research Communications. 2002;**299**(1):126-134

[25] Mitchell D, Bridge R. A test of Chargaff's second rule. Biochemical and Biophysical Research Communications. 2006;**340**(1):90-94

[26] Afreixo V, Bastos CAC, Garcia SP, Rodrigues JMOS, Pinho AJ, Ferreira PJSG. The breakdown of the word symmetry in the human genome. Journal of Theoretical Biology. 2013;**335**: 153-159

[27] Touchon M, Rocha EPC. From GC skews to wavelets: A gentle guide to the analysis of compositional asymmetries in genomic data. Biochimie. 2008; **90**(4):648-659

[28] Mascher M, Schubert I, Scholz U, Friedel S. Patterns of nucleotide asymmetries in plant and animal genomes. Bio Systems. 2013;**111**(3):181-189

[29] Bultrini E, Pizzi E, Del Giudice P, Frontali C. Pentamer vocabularies characterizing introns and intron-like intergenic tracts from *Caenorhabditis elegans* and *Drosophila melanogaster*. Gene. 2003;**304**:183-192

[30] Nikolaou C, Almirantis Y. Deviations from Chargaff's second parity rule in organellar DNA: Insights into the evolution of organellar genomes. Gene. 2006;**381**:34-41

[31] Frank AC, Lobry JR. Asymmetric substitution patterns: A review of possible underlying mutational or selective mechanisms. Gene. 1999;**238**(1):65-77

[32] Guo FB, Yu XJ. Separate base usages of genes located on the leading and lagging strands in *Chlamydia muridarum* revealed by the Z curve method. BMC Genomics. 2007;**8**(1):366

[33] Nikolaou C, Almirantis Y. Mutually symmetric and complementary triplets:

Differences in their use distinguish systematically between coding and non-coding genomic sequences. Journal of Theoretical Biology. 2003;**223**(4): 477-487

[34] Bansal M. DNA structure: Revisiting the Watson-Crick double helix. Current Science. 2003;**85**(11):1556-1563

[35] Mandoiu I, Zelikovsky A. Bioinformatics Algorithms: Techniques and Applications. Vol. 3. John Wiley & Sons; 2008

[36] De Fonzo V, Aluffi-Pentini F, Parisi V. Hidden Markov models in bioinformatics. Current Bioinformatics. 2007;**2**(1):49-61

[37] Niu DK, Lin K, Zhang DY. Strand compositional asymmetries of nuclear DNA in eukaryotes. Jouranl of Molecular Evolution. 2003;**57**(3): 325-334

[38] Qu H, Wu H, Zhang T, Zhang Z, Hu S, Yu J. Nucleotide compositional asymmetry between the leading and lagging strands of eubacterial genomes. Research in Microbiology. 2010;**161**: 838-846