Chloroplast genomes exhibit eight-cluster structuredness and mirror symmetry

Michael Sadovsky[1,2], Maria Senashova[1], and Andrew Malyshev[1]

[1]Institute of computational modelling of SB RAS,
Akademgorodok, 660036 Krasnoyarsk, Russia

[2]Siberian Federal university, Institute of fundamental biology and biotechnology
660049 Russia, Krasnoyarsk, Svobodny prosp., 79

msad@icm.krasn.ru msen@icm.krasn.ru amal@icm.krasn.ru

http://icm.krasn.ru

Abstract. Chloroplast genomes have eight-cluster structuredness, in triplet frequency space. Small fragments of a genome converted into a triplet frequency dictionaries are the elements to be clustered. Typical structure consists of eight clusters: six of them correspond to three different positions of a reading frame shifted for 0, 1 and 2 nucleotides (in two opposing strands), the seventh cluster corresponds to a junk regions of a genome, and the eighth cluster is comprised by the fragments with excessive GC-content bearing specific RNA genes. The structure exhibits a specific symmetry.

Keywords: order, probability, triplet, symmetry, projection, K-means

1 Introduction

Previously, a seven-cluster pattern claiming to be a universal one in bacterial genomes has been reported [1, 2]. This structure was found to be universal, for bacteria; and very elegant theory explaining the observed patterns was proposed. Keeping in mind the most popular theory of chloroplast origin [3–6], we tried to find whether a similar pattern is observed in chloroplast genomes. Surprisingly, eight cluster structure has been found, for chloroplasts, not the seven-cluster one, and the patterns differ rather significantly.

Evidently, such studies are of great evolutionary value: comparing various structures found in DNA sequences of various organisms, one expects to retrieve the evolution process details ranging from races and species to global ecological systems. Here one has to study a three-sided entity: structure, function, and phylogeny. Quite often all three issues are so tightly interweaved that one fails to distinguish the effects and contributions of each issue separately. Here we explore the relation between structure and taxonomy of the bearers of chloroplast genomes. A number of papers aims to study evolutionary processes on the basis of genome sequences structures peculiarities retrieval [8, 7] or a comparative study of some peculiar fragments of genomes [9–14] of chloroplasts.

Let now introduce the strict definitions and exact statements. We shall consider symbol sequences from four-letter alphabet $\aleph = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$ of the length M; the length here is just the total number of symbols (nucleotides) in a sequence. No other symbols or gaps in the sequence take place, by supposition, at least, at the beginning. Any coherent string $\omega = \nu_1\nu_2 \ldots \nu_q$ of the length q makes a word. A structure to be retrieved from chloroplast genomes is provided by clustering of the fragments of equal length isolated within a genome so that each fragment is converted into a triplet frequency dictionary with non-overlapping triplets with no gaps in frame tiling. Thus, we shall keep the consideration within the study of the triplet $\omega_3 = \nu_1\nu_2\nu_3$ frequency dictionaries, only.

Further, we shall consider the genomes of chloroplasts retrieved from EMBL–bank. In case where extra symbols falling out of the alphabet $\aleph$ take place in a sequence, these former were eliminated; the procedure of such elimination is discussed in Subsec. 2.2.

2 Frequency dictionary and genome fragmentation

Indeed, a triplet frequency dictionary could be defined in various ways. The simplest case is provided by the dictionary $W_{(3,1)}$, where the first index shows the length of the words counted in a dictionary, and the second one is the step length (i. e., the number of nucleotides located between two sequential positions of a frame reading). And the frequency dictionary itself is the list of the words (these are the triplets, in our case) found in a sequence, so that each entry of the list is provided by the frequency of that latter. The frequency is defined easily:

$$f_\omega = \frac{n_\omega}{N} \tag{1}$$

where n_ω is the number of copies of the specific word ω, and N is the total number of the counted words (with respect to the copies number);

$$N = \sum_\omega n_\omega \,.$$

For $W_{(3,1)}$ $N = M$, and it is not so in general case. A frequency dictionary W_q of nucleotide sequences is claimed to be an entity bearing a lot of information on that latter [15–20]. A consistent and comprehensive study of frequency dictionaries answers the questions concerning the statistical and information properties of DNA sequences.

In general, one might study a frequency dictionary $W_{(n,m)}$ that comprises the words of the length n counted with the step in m nucleotides. For the purposes of our study, we shall consider the frequency dictionaries $W_{(3,3)}$. Such frequency dictionary is defined ambiguously: there could be three different start positions for triplet counting. Strictly speaking, one should study all three dictionaries of $W_{(3,3)}$ type; moreover, the key issue here is that the three frequency dictionaries $W_{(3,3)}$ differing in the start position exhibit sounding difference in their statistical properties, when determined for coding and non-coding regions of a

genome [1, 2]. This difference yields the clustering standing behind the structuredness we are speaking about.

2.1 Genome fragmentation

For the purposes of the study, we shall not consider all three versions of frequency dictionary $W_{(3,3)}$ differed in start position; on the contrary, we shall define so called *phase* of a fragment. Let now describe the procedure for structuredness retrieval in more detail. Consider a genome sequence that is stipulated to be a symbol sequence from four-letter alphabet $\aleph$. Let then fix the sliding window length L and the step length R figures. Cover then a genome with a tiling windows of the given length moving upright (for definiteness) alongside the sequence, with the step R; if $R < L$ then two windows overlap, otherwise they do not overlap. This is the preliminary transformation of a genome; convert then each identified fragment (of the length L) into the frequency dictionary $W_{(3,3)}$ so that the start position of the reading frame for triplets to coincide to the first nucleotide in the fragment. Thus, a genome is transformed into an ensemble of $W_{(3,3)}$ frequency dictionaries; here each dictionary is labeled with the number of the fragment, as determined alongside the sequence. Finally, we get an ensemble of the points in 63-dimensional metric space, where each point represents a fragment of the genome.

The aim of the work is to reveal the patterns produced by the distribution of those points in 63-dimensional space; formally, the triplet frequencies yield 64-dimensional space, while a triplet must be excluded. Linear constraint

$$\sum_{\omega=\mathsf{AAA}}^{\mathsf{TTT}} f_\omega = 1$$

inflicts rather strong dependence which, in turn, may bring a false signal. Thus, a triplet must be excluded; formally, any triplet may be eliminated. Practically, we excluded the triplet with the lowest standard deviation figure observed over the entire ensemble of frequency dictionaries.

2.2 Fragment phase definition

Previously, three versions of $W_{(3,3)}$ frequency dictionary have been mentioned; they differ in the position of reading frame shift. Here we did not derive all three versions of $W_{(3,3)}$; on contrary, we defined the so called *phase* index for each fragment. The phase is defined by the reciprocal position of a fragment against a coding region. Thus, a fragment is labeled as

phase 0, if the start of a fragment perfectly matches the start of a coding region, or the reminder of the division of the distance from the start position of a coding regions to a fragment by 3 is equal to 0;

phase 1, if the reminder from the division of the distance from the start position of a coding regions to a fragment by 3 is equal to 1;

phase 2, if the reminder from the division of the distance from the start position of a coding regions to a fragment by 3 is equal to 2.

If a part of a fragment falls out of a coding region, the fragment is labeled by *junk phase*. Here we did not distinguish exon-intron structure of a gene.

Actually, the labeling system includes eight items: the phases F_0, F_1 and F_2 correspond to the labels mentioned above, as determined for the leading strand; the phases B_0, B_1 and B_2 correspond to the labels mentions above, as determined for the ladder strand. In this case, the reminder was determined not from the start position of a coding region, but from the end one. Finally, the special phase *tail* was introduced, to identify the peculiar group of fragments within a genome.

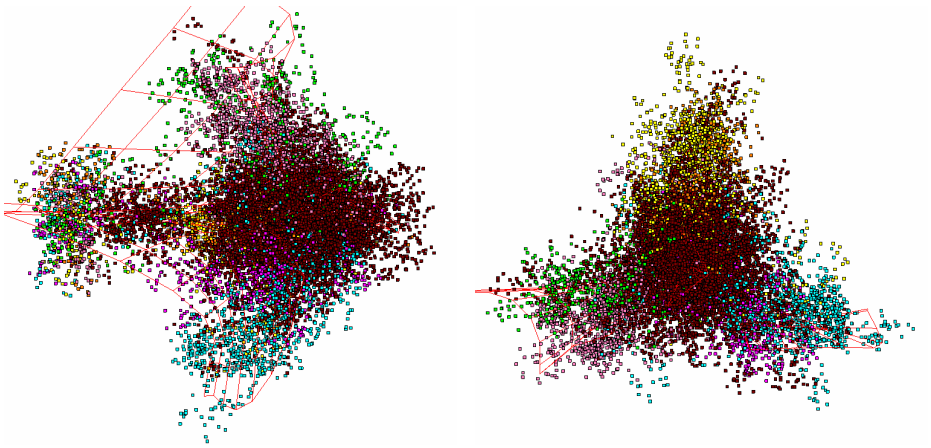
Here the problem of extra symbols arises. Indeed, an elimination of some extras (if any) may cause the shift of the number of a nucleotide position determined alongside the sequence. Such shift may affect the reminder calculation, when a phase is determined. To avoid such deterioration of coding and non-coding regions borders, we remain the numbers of the nucleotides; in other words, an elimination affected both extras, and their numbers in the sequence.

3 Results

We examined 185 chloroplast genomes. Each genome has been covered with a tiling set of fragments, then each fragment has been converted into $W_{(3,3)}$ frequency dictionary, and the phase of each fragment was determined; the dictionaries corresponding to the fragments were marked up with the phase index, as well as with the number of the fragment. We used *ViDaExpert* software [21] to visualize and cluster the data. The greatest majority of genomes exhibits the triplets GCG and CGC having the least standard deviation figures (that were excluded). The excluded triplets form a remarkable couple: they yield the so called *complementary palindrome*, see subsec. 3.2 for details.

The greatest majority of the genomes exhibit similar pattern of the fragments distribution. Fig. 1 shows a typical distribution pattern of the ensemble of fragments converted into $W_{(3,3)}$ frequency dictionaries. This picture presents the chloroplast genome of cranberry *Vaccinium macrocarpon* (AC JQ248601) in EMBL–bank; total length of the genome is 176 037 bp. The length $L = 603$ nucleotides, and $R = 11$ nucleotides. The motivation to fix xuch parameters values is following: L here is comparable to a gene length, and R provides sufficiently dense lattice of a sequence. Definitely, one may choose other parameters figures, while a direct check showed that the pattern is insensitive to them, in rather wide range of parameters.

The points are projected from 63-dimensional Euclidean space determined by triplet frequencies into the three-dimensional Euclidean space determined by three main principal components [22]. The subfigure (a) shows the distribution in "profile" projection (where the first principle component falls on the plane and is directed from left to right; the subfigure (b) shows the same distribution



(a) profile (b) full face

Fig. 1. Cranberry *Vaccinium macrocarpon* chloroplast genome fragments distribution; left is profile view, right is full face view.

in "full face" projection, where the first principle component is normal to the figure plane.

The phases are colored: *phase* F_0 and B_0 are colored in amaranth and cerise, respectively; the *phase* F_1 and B_1 are colored in lemon and orange, respectively; finally, the *phase* F_2 and B_2 are colored in green and cyan, respectively. The *junk* phase is colored in maroon.

Let now concentrate on the left subfigure of Fig. 1. It looks like a kind of fish with a short tail; and the fragments comprising this part of the distribution are those labeled as *tail* phase. The occurrence of this phase differs the chloroplast genomes from bacterial ones. The fragments comprising this *tail* phase are known for its highly increases GC-content value: while the genome-wide figure for that former is 0.38, the specific values for the *tail* phase fragments tends to exceed 0.5 level. As one can see from Fig. 1, the *tail* phase consists of both junk and coding fragments. The *tail* phase fragments present the densely packed cluster of tRNA genes, 16S RNA genes, 23S RNA genes and some other S RNA genes. This cluster has nothing to do with those identified through the mutual distribution of the fragments in the Euclidean space of triplet frequencies. Fig. 2 shows the behaviour of GC-content alongside the genome: *junk* phase is shown in brown, while the coding regions are shown in blue; two very distinct peaks (shown in ovals) in this figure located in diapasons $\sim 110\,000 \leq \sim 115\,000$ and $\sim 165\,000 \leq \sim 170\,000$ comprise the points forming the *tail* phase.

3.1 Clustering vs. visualization

Fig. 1 shows the distribution of the ensemble of $W_{(3,3)}$ frequency dictionaries; so the question arises whether these observations towards the preference in phase lo-

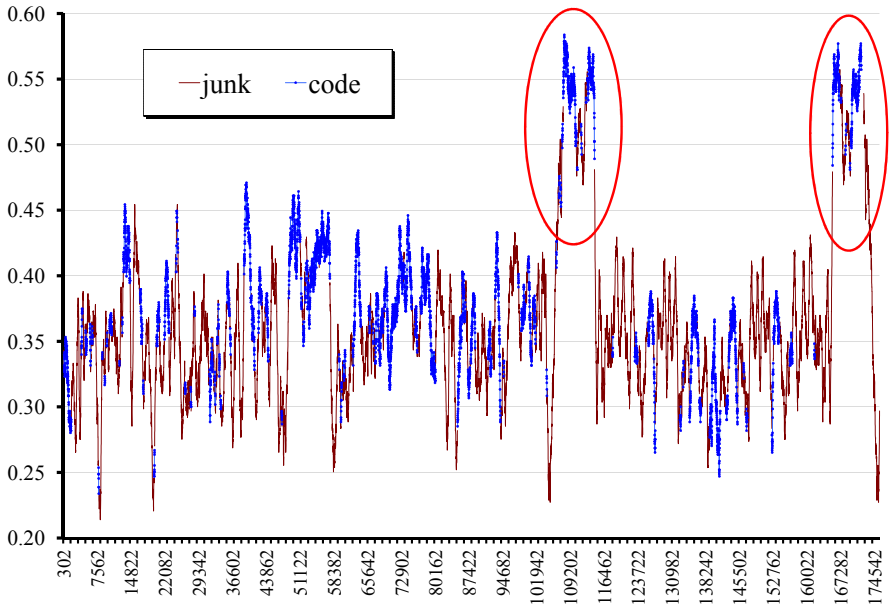
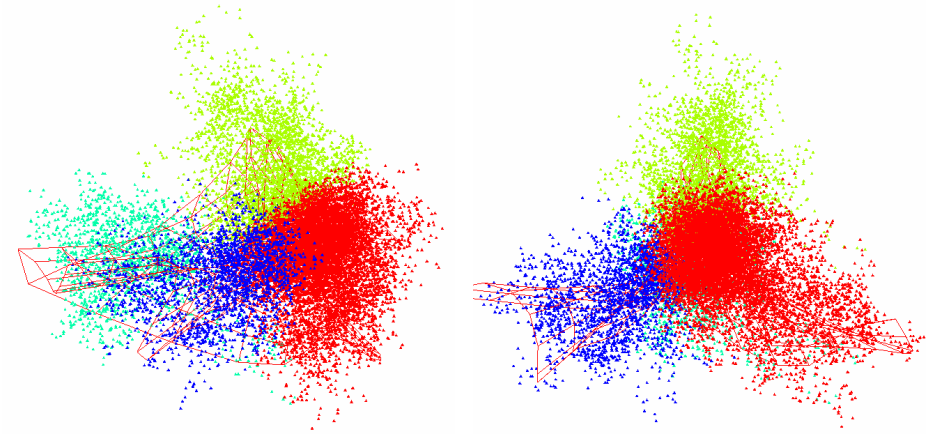


Fig. 2. GC-content of cranberry chloroplast genome determined for each fragment.

cation in the clusters are really existing? In other words, one must check whether a similar clustering could be derived due to some clustering technique. Otherwise, one has to consider the visualized groups of points to be an artifact. To verify it, we have carried out K-means clustering, with $K = 4$ clusters.

K-means clustering yields very stable dispersion of the fragments into four classes. Fig. 3 shows the clustering results. First of all, the clustering is very stable: a hundred of runs of K-means resulted in the same distribution of the points. Next, obviously, K-means for $K = 4$ is unable to dissociate the points of *junk* phase from those belonging to coding regions; an exclusion of the *junk* points from clustering still remains the stable separation into four classes. We did not aim to study clustering of the fragments with an unsupervised cluster technique; on the contrary, the idea was to compare the clusters identified by phases: thus, $K = 4$ seems to be natural, for such test. The test shows good separation, so that the phase defined clusters are not artifacts. Still, the triplet GCG was excluded, for clustering implementation. So, the stability of clustering is proven, hence the beams identified due to visualization are not a artifact, but correspond to naturally determined structure units.



(a) profile (b) full face

Fig. 3. K-means clustering ($K = 4$ of cranberry *Vaccinium macrocarpon* chloroplast genome fragments.

3.2 Symmetry in genome clustering

Let now consider Fig. 1 in more detail. Careful examination of the subfigure (b) shows the specific behaviour of the phases: indeed, the phases $\{F_0, B_0\}$ and $\{F_1, B_1\}$ occupy two opposing beams of the pattern shown in this figure. The phase $\{F_2, B_2\}$ occupy the same beam. This behaviour is not occasional: we have examined 185 chloroplast genomes of ground plants, and all of them exhibit the same phase occupancy.

This behaviour differs completely from similar one observed in bacterial genomes [1, 2]. Two different symmetries stand behind the difference: translational (rotational) symmetry is observed for bacterial genomes, while mirror symmetry is observed for chloroplast ones. The phases patterns (triangles) must be projected one over another, and they are rotated in opposite directions. This symmetry has another manifest in the discrepancy value of the second Chargaff's parity rule determined for centers of those beams. Let's discuss it in more detail.

Chargaff's symmetry of phase clusters The first Chargaff's parity rule stipulates a proximity (rather tight) of the fractions of A's and T's, as well as the fractions of C's and G's counted over a genome. The second Chargaff's parity rule says that the fraction of the strings comprising the *complementary palindrome* are also rather close. That former is a couple of words (of the length q) that are read equally in opposite directions, with respect to the complementarity rule (that was originally formulated for a double strand DNA molecule): G $\leftrightarrow$ C and A $\leftrightarrow$ T. The point is that the fractions (same as frequencies) are counted over a single strand, with neither respect to the second one.

Typical example of a couple of triplets making a complementary palindrome are the triplets that were excluded, when clustering was carried out: GCG $\Leftrightarrow$ CGC; another example is the couple GCCGTAGT $\Leftrightarrow$ ACTACGGC. Two genetic entities could be compared through the discrepancy calculation determined over a frequency dictionary (or two of them):

$$\mu_q = \frac{2}{4^q} \sqrt{\sum_{\omega^* \in \Omega} \left(f_{\omega^*} - f_{\overline{\omega}} \right)^2}, \tag{2}$$

where ω^* and $\overline{\omega}$ are the words comprising a complementary palindrome.

So, the symmetry observed in chloroplast genomes could manifest in (2) figures determined for various beams of the pattern shown in Fig. 1. We calculated μ value (2) for all three beams identified in Fig. 1 (see also Fig. 3), with exclusion of the points belonging to *tail* phase and junk; these values are

$$\mu_1 = 0.001350\,,\ \mu_2 = 0.001224 \quad \text{and} \quad \mu_3 = 0.000290\,.$$

Obviously, the third beam has an order less discrepancy figure, in comparison to two others. These figures have been obtained for 32 couples of triplets of an arithmetic mean of the frequencies of the points of each beam.

Similar pattern is observed for inter-beam discrepancy calculations. Here unlike in (2), one must sum up the squared differences of the frequencies of all 64 couples, since there is no guarantee of the equivalence of two differences

$$f^{(1)}_{\omega} - f^{(2)}_{\overline{\omega}} \qquad \text{and} \qquad f^{(1)}_{\overline{\omega}} - f^{(2)}_{\omega}\,,$$

where the superscript indicates two compared beams. The observed figures are the following:

$$\rho\,(\text{beam}_1, \text{beam}_2) = 0.011991\,, \quad \rho\,(\text{beam}_1, \text{beam}_3) = 0.051165\,,$$
$$\rho\,(\text{beam}_3, \text{beam}_2) = 0.054165\,.$$

Again, the beam # 3 is isolated from two others. The direct comparison of the means and the clusters comprised from various phases unambiguously proves that the beam # 3 is the cluster consisting of F_2 and B_2 phases.

4 Discussion

The labeling system of the formally identified fragments in a sequence seems to be rather strict and to provide a kind of bias in favor of the non-coding regions. A rough estimation of the number of border fragments (i. e., those that fall both in coding and non-coding regions of a genome) in the ensemble is small enough. Suppose, the number of coding regions in a chloroplast genome is 50. Then the approximate number of such border fragments is about $L \times R^{-1} \times 50 \approx 3000$. This estimation shows rather significant bias to *junk* labeled fragments resulted from the border fragments, so it may deteriorate the patterns from these border

fragments. Yet, further investigation is necessary to answer this question, while the hypothesis is that the impact of those border fragment is not significant, currently.

In papers [1, 2] an approach to reveal a structuredness in bacterial genomes based on the comparison of frequency dictionaries $W_{(3,3)}$ of the fragments of a genome is presented; our results show that chloroplasts behave in other way always clustering in two coinciding triangles. The vertices of that latter correspond to phases of a reading frame comprising the fragments with identical reading frame shift figure (reminder value). Another important issue is that GC-content does not determine the positioning of the clusters, unlike for bacterial genomes.

A mirror symmetry in frequency dictionaries of $W_{(3,3)}$ type is the most intriguing issue of the work: such symmetry was never observed in bacterial genomes, nor in yeast genomes, nor in the genomes of some other higher organisms. Whether this mirror symmetry is the specific feature of chloroplasts, or it is peculiar for any organelle, is a matter of question. For chloroplasts, the symmetry has been checked for a number of genomes of the plants of various taxa. An idea to reveal some similarities in the patterns described above, in chloroplast genomes, and in some other genetic entities which are claimed to be a kind of relatives of chloroplasts was disproved: we checked several cyanobacteria genomes for the pattern occurrence, and nothing similar was found [23].

Careful examination of the databases implemented for each studied genome (see also Fig. 1) shows a relative maintenance of the fraction of the fragments labeled F_k and B_k; indeed, the set of genomes could be separated into two subsets: the former with $n_{F_k} > n_{B_k}$, and the latter with $n_{F_k} < n_{B_k}$. Here n_{F_k} (n_{B_k}, respectively) is the fraction of the fragments labeled as n_{F_k}-phase (n_{B_k}-phase, respectively). We hypothesize that the minimum standard deviation triplet (whether it would be GCG or CGC) is determined by the ratio of n_{F_k} and n_{B_k} figures.

Meanwhile, the most exciting observation towards the symmetry in chloroplast genomes consists in the mirror symmetry of the phase-determined clusters comprising the relevant fragments of a genome. Such symmetry manifests also in another type of symmetric-like relation that is expressed in terms of Chargaff's parity rule: the phases F_2 and B_2 always cohere into a single cluster, that is also identified as is by K-means. A verification of this clustering pattern over a number of chloroplast genomes allows to say that these are F_2 and B_2 phases that fall into the same cluster. Moreover, the location of other phases is determined unambiguously, against these two ones. This fact may provide an extremely fast technique for a primary annotation of a *de novo* assembled chloroplast genome: slicing a sequence into an ensemble of the fragments as it is described above and clustering them takes seconds, and reveals the fragments which are almost for sure ascribed to the phase (if the hypothesis on the interrelation between the least standard deviation figure of a triplet, and the ratio of the phase differing fragments holds true), and, what is more important, the fragments are labeled with the reading frame shift figure.

In conclusion, we outline few issues falling beyond the scope of this paper, while expecting an urgent research. The first issue is the study of the chloroplasts from other species that tend to show a deviation from the described pattern (mosses, equisetum, unicellular green algae, etc.). The second issue is more detailed study of the part of genomes that comprise the *tail* phase. Finally, the third issue is the study of "dark matter" of a genome: the fragments that correspond to non-coding regions. Some preliminary investigations show that these fragments also make various structures, and are sensitive to the taxonomy of the bearers of the genomes. More detailed discussion of these issues falls beyond the scope of this paper.

5 Acknowledgement

This study was supported by a research grant # 14.Y26.31.0004 from the Government of the Russian Federation.

References

1. Gorban A. N., Zinovyev A. Yu., Popova T. G. 2003. Seven clusters in genomic triplet distributions. *Silico Biology* **3**(4): 471–482.
2. Gorban A. N., Zinovyev A. Yu., Popova T. G. 2005. Four basic symmetry types in the universal 7-cluster structure of microbial genomic sequences. *Silico Biology.* **5**(3): 265–282.
3. Mereschkowsky K.S. 1910. Theorie der zwei Plasmaarten als Grundlage der Symbiogenesis, einer neuen Lehre von der Entstehung der Organismen. *Biol Centralbl.* **30**: 353–367.
4. Mereschkovsky K.S. 1905. Über Natur und Ursprung der Chromatophoren im Pflanzenreiche. *Biol. Zentr.-Bl. Bd.* **85**(18): 593–604.
5. Zimorski V., Ku Ch., Martin W. F., Gould S. B. 2014. Endosymbiotic theory for organelle origins *Current Opinion in Microbiology* **22**: 38–48.
6. Raven J. A., Allen J. F. 2003. Genomics and chloroplast evolution: what did cyanobacteria do for plants? *Genome Biology.* **4**(3): 209.
7. Carbonell-Caballero J., Alonso R., Ibanez V., Terol J., Talon M., Dopazo J. 2015. A phylogenetic analysis of 34 chloroplast genomes elucidates the relationships between wild and domestic species within the genus citrus *Mol. Biol. Evol.* **32**(8): 2015–2035.
8. Leliaert F., Smith D.R., Moreau H., Herron M.D., Verbruggen H., Delwiche Ch.F., De Clerck O. 2012. Phylogeny and Molecular Evolution of the Green Algae *Critical Reviews in Plant Sciences* **31**: 1–46.
9. Katayama H., Ogihara Y. 1996. Phylogenetic affinities of the grasses to other monocotsas revealed by molecular analysis of chloroplast DNA *Current Genetics* **29**: 572–581.
10. Milanowski R., Zakrys B., Kwiatowski J. 2001 Phylogenetic analysis of chloroplast small subunit rRNA genes of the genus Euglena Ehrenberg *Int. J. of Systematic and Evolutionary Microb.* **51**: 773–781.

11. Marazzi B., Endress P.K., De Queiroz L.P., Conti E. 2006. Phylogenetic relationships within senna (leguminosae, cassiinae) based on three chloroplast DNA regions: patterns in the evolution of floral symmetry and extrafloral nectaries *Am. J. of Botany.* **93**(2): 288–303.
12. Shaw J., Lickey E.B., Beck J.T., Farmer S.B., Liu W., Miller J., et al. 2005. The tortoise and the hare ii: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis *Am. J. of Botany* **92**(1): 142–166.
13. Dong W., Liu J., Yu J., Wang L., Zhou Sh. 2012. Highly Variable Chloroplast Markers for Evaluating Plant Phylogeny at Low Taxonomic Levels and for DNA Barcoding *PLoS ONE* **7**(4): 1–9.
14. Gielly L., Taberlet P. 1994. The Use of Chloroplast DNA to Resolve Plant Phylogenies: Noncoding versus rbcL Sequences *Mol. Biol. and Evolution* **11**(5): 769–777.
15. 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**: 265–278.
16. Gorban A. N., Popova T. G., Sadovsky M. G., Wünsch D. C. 2001. Information content of the frequency dictionaries, reconstruction, transformation and classification of dictionaries and genetic texts. *Intelligent Engineering Systems through Artificial Neural Networks*, 11 — *Smart Engineering System Design*, N.-Y.: ASME Press, (2001). pp.657 – 663.
17. Gorban A. N., Popova T. G., Sadovsky M. G. 2000. Classification of symbol sequences over thier frequency dictionaries: towards the connection between structure and natural taxonomy. *Open Systems & Information Dyn.* **7**: 1–17.
18. Sadovsky M. G., Shchepanovsky A. S., Putintzeva Yu. A. 2008. Genes, Information and Sense: Complexity and Knowledge Retrieval. *Theory in Biosciences* **127**: 69–78.
19. Sadovsky M. G. 2003. Comparison of real frequencies of strings vs. the expected ones reveals the information capacity of macromoleculae. *J.of Biol.Physics* **29**: 23–38.
20. Sadovsky M. G. 2006. Information capacity of nucleotide sequences and its applications. *Bulletin of Math.Biology.* **68**: 156–178.
21. *http://bioinfo-out.curie.fr/projects/vidaexpert/*
22. Gorban A. N., Zinovyev A. Yu. 2010. Principal manifolds and graphs in practice: from molecular biology to dynamical systems. *Int. J. of Neural Systems*, **20**(3): 219–232.
23. Sadovsky M. G., Senashova M. Yu., Malyshev A. V. 2018 Eight cluster structuredness of genomes of ground plants. *Russian J. of Gen.Biol.*, **79**(2): in press