EIGHT-CLUSTER STRUCTURE OF CHLOROPLAST GENOMES DIFFERS FROM SIMILAR ONE OBSERVED FOR BACTERIA

Michael G.Sadovsky, 1, 2 Maria Yu. Senashova, 1 and Andrew V. Malyshev 1

1 Institute of computational modelling of SD of RAS; 660036 Russia, Krasnoyarsk, Akademgorodok. 1
2 Siberian Federal university; 660041 Russia, Krasnoyarsk, Svobodny prosp. 79.

Previously, a seven-cluster pattern claiming to be a universal one in bacterial genomes has been reported. Keeping in mind the most popular theory of chloroplast origin, we checked whether a similar pattern is observed in chloroplast genomes. Surprisingly, eight cluster structure has been found, for chloroplasts. The pattern observed for chloroplasts differs rather significantly, from bacterial one, and from that latter observed for cyanobacteria. The structure is provided by clustering of the fragments of equal length isolated within a genome so that each fragment is converted in triplet frequency dictionary with non-overlapping triplets with no gaps in frame tiling. The points in 63-dimensional space were clustered due to elastic map technique. The eight cluster found in chloroplasts comprises the fragments of a genome bearing tRNA genes and exhibiting excessively high GC-content, in comparison to the entire genome.

PACS numbers: 87.10.+e, 87.14.Gg, 87.15.Cc, 02.50.-r

I. INTRODUCTION

Detailed study of a structure of nucleotide sequences is a key issue in up-to-date molecular biology and bioinformatics. Such studies are carried out in two (interconnecting) paradigms: the former is structure-function relationship, and the latter is evolutionary one. A retrieval of the interrelation between structure and function of various biological macromolecules is a core issue of up-to-date molecular and system biology. Currently, a huge number of publications appears annually on this subject; yet, the problem is still far from any completion. Moreover, some new structures are reported nowadays 1, 2.

Evolutionary value of such studies is rather apparent: 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. Meanwhile, such studies face a number of problems in selection and quality of biological material to be considered. Skipping off possible errors in sequencing and/or annotation of genetic entities, one faces a great complexity of genomes, or even separate chromosomes. Here one has to study a three-sided entity: structure, function, and phylogeny. Quite often all three issues are so tightly interwoven that one fails to distinguish the effects and contributions of each issue separately.

Prokaryotic organisms seem to be rather suitable for this type of researches: bacterial genome is considerably short and always consists of a single chromosome. An ambiguity in bacterial taxonomy looks like a pay-off for the genome simplicity of these organisms; the problem in taxonomy grows up, as higher taxa are considered 3, 4.

In such capacity, organella genomes seem to be very suitable object for the studies of the type mentioned above: keeping oneself within the organella of the same type (say, chloroplasts), one avoids any problems related to a diversity of functional charge of a genome.

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 5, 6 or a comparative study of some peculiar fragments of genomes 1, 2, 7–10 of chloroplasts.

Sounding diversity of structures that could be found in DNA sequences is another problem. Surely, the problem hardness depends on the notion of a structure in DNA molecule. Hereafter a structure is stipulated to be a pattern in mutual interlocation of nucleotides manifesting in statistical properties of formally identified short fragments of a genome, i.e. the ensemble of strings of the given length q. Further we shall concentrate on the ensembles of strings of the length q = 3 (triplets). Henceforth, the list of triplets observed within a genome or its part accompanied with the frequencies of these former is the structure under consideration; see details below.

Indeed, we shall concentrate on the study of mutual location of the points of 63-dimensional space of triplet frequencies, where each point corresponds to a fragment identified within a genome, due to some regular procedure. The matter of interest is a cluster structuredness (if any) of those fragments of genomes converted into frequency dictionaries of triplets, in 63-dimensional metric space. Such approach has been originally explored by Alexander Gorban and co-authors 11, 12, for bacterial genomes. They have found seven-cluster patterns in the fragments distribution, where the specific type of the pattern is strongly ruled by GC-content of a genome.

The most popular theory of chloroplast origin, that is bacterial symbiogenesis theory 13, 15, stipulates a
relation between some bacteria, and chloroplasts; that is the motivation standing behind our study: whether this relation manifests in a similarity of the patterns observed for bacteria [11, 12], and those observed for chloroplasts. Briefly speaking, the answer is negative. Moreover, chloroplast genomes exhibit rather specific patterns drastically differing them from bacterial genomes, and further we shall present the result demonstrating the difference, and discuss this point.

II. MATERIALS AND METHODS

Chloroplast genomes were retrieved from EMBL–bank; the list of entities comprises 178 entries. Table I enlists the genomes under consideration. Evidently, genomes differ in the quality of sequencing: they may have extra symbols other than $\mathcal{N} = \{A, C, G, T\}$. Wherever it happened, we eliminated such extra symbols from a sequence, and concatenated it into the coherent one.

A. Frequency dictionaries and genome fragmentation

We stipulate a genome to be a coherent sequence from four-letter alphabet $\mathcal{N} = \{A, C, G, T\}$: the number $N$ of nucleotides is the length of the sequence. Let then fix the length $q$ of a window, and the length $t$ of a step. Moving then the window upright (for certainty) alongside the sequence with the step $t$ and counting the number $n_\omega$ of strings $\omega$ of the given length $q$ identified by the sequential locations of the window, one gets the finite dictionary $F_{(q,t)}$. Changing the numbers $n_\omega$ for frequencies

$$f_\omega = \frac{n_\omega}{M} \quad \text{with} \quad M = \sum_\omega n_\omega,$$  

one gets the frequency dictionary $W_{(q,t)}$: $M = N$ for $W_{(q,1)}$. Actually, such definition of $W_{(q,t)}$ requires to connect the sequence under consideration into a ring (see details in [11 12 19]).

Everywhere further we shall constrain with the dictionaries $W_{(3,3)}$, only. It enlists the triplets counted with no overlapping, while with no gaps between the frames. The choice of $q = 3$ and $t = 3$ is motivated by apparent biological issues: triplets yield the strongest signal, in DNA sequences, and the step $t = 3$ reveals it, that is coding positions in DNA sequence.

The frequency dictionary $W_{(3,3)}$ exists in three different (in general) versions differing in reading frame position; that latter is called phase of a fragment below. Indeed, for the sequence

$$\text{CTTGTCGATCAATTGTGATCCAGTTTTATGATTGCACCGCAGAAAGTG}$$

the dictionaries $W_{(3,3)}$: $W_{(3,3)}^{(0)}$, $W_{(3,3)}^{(1)}$ and $W_{(3,3)}^{(2)}$ are shown in Table I. Here the number of copies, not the frequencies are shown. Everywhere below, we shall describe the frequency dictionaries $W_{(3,3)}^{(0)}$, $W_{(3,3)}^{(1)}$ and $W_{(3,3)}^{(2)}$ for the sequence (2).

| $W_{(3,3)}^{(0)}$ | $W_{(3,3)}^{(1)}$ |
|-----------------|-----------------|
| $W_{(3,3)}^{(2)}$ | $W_{(3,3)}^{(1)}$ |
| $W_{(3,3)}^{(0)}$ | $W_{(3,3)}^{(1)}$ |
| $W_{(3,3)}^{(2)}$ | $W_{(3,3)}^{(1)}$ |

TABLE I: The dictionaries $W_{(3,3)}^{(0)}$, $W_{(3,3)}^{(1)}$, and $W_{(3,3)}^{(2)}$ for the sequence (2).

velop the frequency dictionaries $W_{(3,3)}^{(0)}$, for each fragment of a sequence. Reciprocally, the phase of the fragment has been determined, instead of the implementation of two other dictionaries; see details below.

To figure out the inner structuredness of a chloroplast genome, we cut it into a set of (overlapping) fragments. To do that, the length of a fragment $L$ and the move step $R$ alongside a genomes have been fixed; we used the figures $L = 603$ and $R = 11$, in our studies. The motivation for the choice of such figures is following: we need to choose the length $L$ of a fragment to be odd and divisible by 3, while the step $R$ must be not divisible by 3. Next, the length of a fragment is chosen rather close to a gene length. The step length $R$ determines the number of points taken into consideration, e.g. for $K$-means clustering; the chosen step figure yields $\sim 10^4$ fragments (later converted into the points in a metric space). Obviously, both $L$ and $R$ could be set de novo, if necessary.

Any frequency dictionary $W_{(3,3)}$ maps a sequence into a point in 63-dimensional space. Indeed, the total number of triplets is equal to 64; meanwhile, the linear constraint

$$\sum_\omega n_\omega = 1$$

makes remain only 63 ones independent; the frequency
of the last one is unambiguously determined from \( [3] \). Formally, any triplet may be eliminated; practically, we excluded the triplet exhibiting the least standard deviation figure determined over the entire ensemble of the fragments. Table [I] shows these triplets, at the column labeled \( \omega_{\text{min}} \).

Apparently, there might be other ways to determine the excluded triplet. For example, it is useful to exclude the variable with maximum value, for some situations; here we followed the described above way, since the least standard deviation of a triplet frequencies observed over a dataset means the least distinguishability of the objects comprising a dataset, over this variable. Thus, the dimensionality of the space to cluster the frequency dictionaries of triplets becomes equal to 63.

1. The phase of a fragment of sequence

Previously, three types of frequency dictionaries \( W^{(0)}_{(3,3)}, W^{(1)}_{(3,3)} \) and \( W^{(2)}_{(3,3)} \) were shown (see Table [I]). Meanwhile, we developed only one frequency dictionary; that was \( W^{(0)}_{(3,3)} \) dictionary. The fragment was then labeled using one of four labels: phase 1, phase 2, phase 3 and junk. The label was determined by the location of a fragment within a sequence; to do that, we used the annotation of each genome under consideration.

A fragment was labeled as

- **junk**, if it contains at least a half of a non-coding region of a genome within itself;
- **phase 0**, if the center of a fragment falls into a coding region of a genome, and the reminder of the division of the distance between the central nucleotide of a fragment, and the starting nucleotide of a coding region is equal to 0;
- **phase 1**, if the center of a fragment falls into a coding region of a genome, and the reminder of the division of the distance between the central nucleotide of a fragment, and the starting nucleotide of a coding region is equal to 1;
- **phase 2**, if the center of a fragment falls into a coding region of a genome, and the reminder of the division of the distance between the central nucleotide of a fragment, and the starting nucleotide of a coding region is equal to 2.

For genes (or coding regions) located in the ladder strand, the above mentioned procedure still holds true, but the distance to the central nucleotide of a fragment is determined not from the start position (formally indicated in a file), but from the end of that latter.

B. Clustering of frequency dictionaries

As soon, as the fragments are converted into the frequency dictionaries \( W^{(0)}_{(3,3)}, \) then each dictionary was labeled with the number of the nucleotide occupying the central position at the corresponding fragment. Also, each fragment was labeled with its phase. To make the space of frequency dictionaries metric, one must implement a metrics; there is a number of options here (see [21, 28] for details). Meanwhile, we use Euclidean metrics:

\[
\rho(W^{(0)}_{(3,3)}, W^{(1)}_{(3,3)}) = \sqrt{\frac{1}{\omega=AAA} \sum_{\omega=AAA} (f_\omega^{(1)} - f_\omega^{(2)})^2}\ . \quad (4)
\]

Here \( f_\omega^{(2)} \) is the frequency of a triplet \( \omega \) observed in the \( j \)th frequency dictionary; this index has nothing to do with the frame shift described above.

We studied the distribution of those fragments, in 63-dimensional space using VidaExpert (http://bioinfo-out.curie.fr/projects/vidaexpert/) software. No special technique for clustering has been used: we identified the clusters as is, through visualization. Nonetheless, all the clusters identified through visualization were also identified with K-means; thus, those clusters could be verified objectively.

In addition, GC-content has been determined, both for each fragment, and the genome entirely (see Table [I] GC labeled column).

III. RESULTS

First, let’s consider the list of chloroplast genomes used in the study, in more detail. The list is quite homogeneous, in terms of the length of sequences; thus, we may not expect any effect resulted from a length difference. Next point is the eliminated triplet choice; Table [II] shows those triplets at the sixth column. Actually, there are only four triplets eliminated in various genomes: CGC (58 entries), GCG (113 entries), GAC (1 entry) and TAA (also 1 entry).

The triplets CGC and GCG are of great interest: they both are palindromes (read equally in opposite directions), and besides they together comprise the couple of so called *complementary palindrome*. That latter consists of two string (triplets, in our case) that are read equally in opposite directions, with respect to Chargaff’s parity rule: CGC ⇔ GCG. Such symmetry is rather important both in analysis, and in biological issues standing behind it; more detailed discussion see below.
TABLE II: List of genomes studied; \(N\) is the length of genome, \(\text{GC}\) is GC-content, \(J\) is junk percentage, and \(\omega_{\text{min}}\) is the triplet with minimal standard deviation.

| Genomes                        | AC number | \(N\)     | \(\text{GC}\) | \(J\)     | \(\omega_{\text{min}}\) |
|--------------------------------|-----------|-----------|--------------|-----------|--------------------------|
| Allium cepa (onion)            | KF728080  | 153538    | 0.37         | 41.84     | CCG                      |
| Anemone mirabilis              | EU043314  | 108097    | 0.41         | 56.11     | CCG                      |
| Angiosperma erecta             | DQ821119  | 153901    | 0.35         | 54.15     | CCG                      |
| Anthoceros angustus            | AB086179  | 161162    | 0.33         | 48.48     | CCG                      |
| Apollonia endhiviiolata         | JX827163  | 120544    | 0.36         | 38.62     | CCG                      |
| Arabidopsis thaliana           | AP000423  | 154478    | 0.36         | 48.76     | CCG                      |
| Brachypodium distachyon         | EU325680  | 135197    | 0.39         | 69.56     | CCG                      |
| Cyclas revoluta                | JN667388  | 162489    | 0.39         | 45.33     | CCG                      |
| Equisetum arvense              | GU191334  | 133093    | 0.33         | 45.07     | CCG                      |
| Fagopyrum esculentum            | EF425477  | 139596    | 0.38         | 47.91     | CCG                      |
| Fargesia nitida                | JX513416  | 139535    | 0.39         | 51.45     | CCG                      |
| Fragaria chiloensis             | JX884816  | 156603    | 0.37         | 48.83     | CCG                      |
| Fritillaria napelhensis         | KP712486  | 152145    | 0.37         | 52.61     | CCG                      |
| Galolognshania megalothyrsia    | JX513419  | 140064    | 0.39         | 57.58     | CCG                      |
| Genista marginalis              | HG550134  | 141252    | 0.38         | 58.61     | CCG                      |
| Ginkgo biloba                   | AB684440  | 156945    | 0.40         | 53.00     | CCG                      |
| Glycine max                     | DQ315723  | 152218    | 0.35         | 48.49     | CCG                      |
| Glycyrrhiza glabra              | KF201590  | 127942    | 0.34         | 47.78     | CCG                      |
| Gnetum montanum                 | KC427277  | 115091    | 0.38         | 46.70     | CCG                      |
| Goodyera fumata                 | KJ509199  | 155563    | 0.37         | 48.37     | CCG                      |
| Gossypium anomalum              | JF317351  | 159505    | 0.37         | 50.02     | CCG                      |
| Guizotia abyssinica             | EU549769  | 151762    | 0.38         | 48.47     | CCG                      |
| Habanaria pantlingana           | JX521048  | 153951    | 0.37         | 48.25     | CCG                      |
| Helianthus annuus               | DQ838315  | 151104    | 0.38         | 48.55     | CCG                      |
| Hibiscus syriacus               | KP688069  | 161019    | 0.37         | 60.43     | CCG                      |
| Hordeum vulgare                 | KC912688  | 114434    | 0.37         | 50.92     | CCG                      |
| Hyperzia lucidula               | AY660566  | 154373    | 0.36         | 52.05     | CCG                      |
| Hyoscyamus niger                | KF248009  | 155720    | 0.38         | 48.91     | CCG                      |
| Hypsocharis bilobata            | KF240616  | 165002    | 0.39         | 49.61     | CCG                      |
| Illicium oligandrum             | EF380354  | 148553    | 0.39         | 52.06     | CCG                      |
| Indosasa sinica                 | JX513422  | 139660    | 0.39         | 57.70     | CCG                      |
| Iochroma nitidum                | KP294386  | 156574    | 0.38         | 50.61     | CCG                      |
| Ipomoea batatas                 | KP212149  | 161303    | 0.38         | 53.18     | CCG                      |
| Jacobaea vulgaris               | HG254689  | 150688    | 0.37         | 49.17     | CCG                      |
| Jasminum nudiflorum             | DQ673255  | 165121    | 0.38         | 49.87     | CCG                      |
| Jatropha curcas                  | FJ695500  | 163856    | 0.35         | 51.33     | CCG                      |
| Juniperus bermudiana            | KF866297  | 127631    | 0.35         | 40.85     | CCG                      |
| Juniperus monosperma            | KF866298  | 127744    | 0.35         | 40.75     | CCG                      |
| Juniperus virginiana            | KF866300  | 127770    | 0.35         | 40.92     | CCG                      |
| Kalopanax septemlobus           | KC456167  | 156413    | 0.38         | 50.68     | CCG                      |
| Keteleeria davidiana            | AP010820  | 117720    | 0.39         | 45.19     | CCG                      |
| Lactuca sativa                  | AP007232  | 152765    | 0.38         | 52.30     | CCG                      |
| Larix decidua                   | AB501189  | 122474    | 0.39         | 45.12     | CCG                      |
| Lathyrus sativus                | HM029371  | 121020    | 0.35         | 44.25     | CCG                      |
| Lecontella madagascariensis      | HF543599  | 139073    | 0.39         | 57.13     | CCG                      |
| Lemma minor                     | DQ400350  | 165955    | 0.36         | 49.42     | CCG                      |
| Lepidium virginicum             | AP009374  | 154743    | 0.36         | 48.68     | CCG                      |
| Licania alba                     | KJ414483  | 162467    | 0.36         | 51.53     | CCG                      |
| Lindenbergia philippensis       | HG530133  | 155103    | 0.38         | 48.93     | CCG                      |
| Liquidambar formosana           | KC588388  | 160410    | 0.38         | 50.85     | CCG                      |
| Liriodendron tulipifera         | DQ899947  | 159886    | 0.39         | 50.29     | CCG                      |
| Lobularia maritima              | AP009375  | 152659    | 0.37         | 48.37     | CCG                      |
| Lolium perenne                   | AM777385  | 135282    | 0.38         | 55.66     | CCG                      |
| Loniceria japonica              | KJ70923   | 155078    | 0.39         | 51.72     | CCG                      |
| Lotus japonicus                 | AP002983  | 150519    | 0.36         | 48.85     | CCG                      |
| Lupinus luteus                   | KC695666  | 151891    | 0.37         | 48.73     | CCG                      |

continued on the next page
| Genomes                           | AC number     | N    | GC | J  | $\omega_{\text{min}}$ |
|----------------------------------|---------------|------|----|----|----------------------|
| Lygodium japonicum               | KC536645      | 157260 | 0.41 | 48.69 | GCC U                |
| Magnolia kwangsiensis            | HM775382      | 159667 | 0.39 | 50.92 | GCC U                |
| Manihot esculenta                | EU117376      | 161453 | 0.36 | 55.23 | GCC U                |
| Mankyuca chejuensis              | JF343520      | 146221 | 0.38 | 52.36 | GCC U                |
| Marchantia paleacea              | X04465        | 121024 | 0.29 | 39.93 | GCG D                |
| Marsilea crenata                 | KC536646      | 151628 | 0.42 | 47.23 | GCC U                |
| Masdevallia cocinea              | KP205432      | 157423 | 0.37 | 49.71 | GCC U                |
| Megaleranthus saniculifolia      | FJ597983      | 159924 | 0.38 | 50.77 | GCC U                |
| Metapanax delavayi               | KC456165      | 156343 | 0.38 | 50.65 | GCC U                |
| Millettia pinnata                | JN673818      | 152968 | 0.35 | 49.03 | GCC U                |
| Morus indica                     | DQ226511      | 158484 | 0.36 | 50.05 | GCG U                |
| Myriopteris lindheimeri          | HM778032      | 155770 | 0.43 | 48.17 | GCG U                |
| Nageia nagi                      | AB830885      | 133722 | 0.37 | 44.15 | GCG D                |
| Najas flexilis                   | JX978472      | 156362 | 0.38 | 58.16 | GCC D                |
| Nasturtium officinale            | AP009376      | 155105 | 0.36 | 49.00 | GCC U                |
| Nelumbo lutea                    | FJ754269      | 163206 | 0.38 | 51.25 | GCC U                |
| Neyraudia reynaudiana            | KF356392      | 133567 | 0.38 | 55.66 | GCC U                |
| Nicotiana sylvestris             | AB237912      | 155941 | 0.38 | 46.15 | GCC U                |
| Nuphar advena                    | DQ354691      | 160866 | 0.39 | 50.64 | GCC U                |
| Nymphaea alba                    | AJ627251      | 159930 | 0.39 | 49.50 | GCC U                |
| Oenothera argillicola            | EU262887      | 165061 | 0.39 | 50.34 | GCC U                |
| Olea europaea                    | GU228899      | 155888 | 0.38 | 49.06 | GCC U                |
| Oligostachyum shiuyingianum      | JX513423      | 139647 | 0.39 | 57.64 | GCC U                |
| Olmarabidopsis pumila            | AP009368      | 154737 | 0.36 | 49.16 | GCC U                |
| Ophioglossum californicum        | KC117178      | 138270 | 0.42 | 47.57 | GCC U                |
| Orobanchis gracilis              | HG803179      | 65533  | 0.35 | 73.30 | GCC U                |
| Orthotrichum rogeri              | KP119739      | 123363 | 0.28 | 43.01 | GCC D                |
| Oryza nivara                     | JP861109      | 134448 | 0.39 | 54.61 | GCC D                |
| Oryza sativa                     | JN861110      | 134449 | 0.39 | 50.23 | GCC D                |
| Oryza sativa Indica              | JN861110      | 134459 | 0.39 | 55.21 | GCC D                |
| Oryza sativa Japonica            | JN522329      | 134496 | 0.39 | 63.88 | GCC D                |
| Oryza sativa Japonica            | JN522329      | 134551 | 0.39 | 66.96 | GCC D                |
| Oryza sativa Japonica            | JN522329      | 134551 | 0.39 | 64.36 | GCC D                |
| Oryza sativa Japonica            | JN522329      | 134551 | 0.39 | 42.96 | GCC D                |
| Pachycladon cheesemani           | JQ806762      | 154498 | 0.36 | 49.20 | GCC U                |
| Paeonia obovata                  | KJ206533      | 152696 | 0.38 | 49.68 | GCC U                |
| Panax ginseng                    | AJ582139      | 156313 | 0.38 | 49.32 | GCC U                |
| Panicum virgatum                | HQ731441      | 139677 | 0.39 | 57.05 | GCC U                |
| Paphiopedilum armeniacum         | KJ566307      | 162682 | 0.35 | 55.30 | GCC D                |
| Parinari campestris              | KJ414486      | 162637 | 0.36 | 51.58 | GCC U                |
| Parthenium argentatum            | GU120098      | 152803 | 0.38 | 77.21 | GCC D                |
| Pelargonium × hortorum           | DQ897081      | 217942 | 0.40 | 48.14 | GCC D                |
| Pentactina rugicola              | JQ417663      | 166612 | 0.37 | 49.40 | GCC D                |
| Penthorum chinense               | JX436155      | 156686 | 0.37 | 51.26 | GCC U                |
| Phalaenopsis equestris           | JF719062      | 148959 | 0.37 | 56.71 | GCC U                |
| Pharus lappulaceus               | KC311467      | 141928 | 0.38 | 58.51 | GCC U                |
| Phoenix dactylifera              | GU811709      | 158462 | 0.37 | 50.00 | GCC U                |
| Phragmites australis             | KP730315      | 137561 | 0.39 | 56.39 | GCC U                |
| Thyllostachys edulis             | HQ337796      | 139679 | 0.39 | 57.14 | GCC D                |
| Thyllostachys propinqua          | JN415113      | 139704 | 0.39 | 57.54 | GCC D                |
| Thyssosmitrella patens           | AP009367      | 122890 | 0.29 | 42.08 | GCC U                |
| Ticea abies                      | HF937082      | 124084 | 0.39 | 51.11 | GCC U                |
| Yucca elata                      | HG803178      | 147140 | 0.38 | 56.14 | GCC U                |
| Punus contorta                   | EU998740      | 115267 | 0.38 | 49.93 | GCC U                |
| Punus taeda (lobolly pine)       | KC427273      | 121530 | 0.39 | 44.26 | GCC D                |
| Piper cenocladium                | DQ887677      | 160624 | 0.38 | 52.93 | GCC U                |
| Pismum sativum                   | HM029370      | 122169 | 0.35 | 46.08 | GCC U                |
| Pleioblastus maculatus           | JX513424      | 139720 | 0.39 | 57.73 | GCC D                |

continued on the next page
| Genomes                  | AC number | N   | GC | J   | $\omega_{\text{min}}$ |
|-------------------------|-----------|-----|----|-----|-----------------------|
| Podocarpus lambertii    | KJ010812  | 1337 | 34 | 44  | 25                   |
| Populus alba            | AP008956  | 1565 | 34 | 49  | 21                   |
| Premna microphylla      | KM981744  | 1552 | 34 | 48  | 79                   |
| Primula poissonii       | KF753634  | 1516 | 34 | 47  | 61                   |
| Prunus kansuensis       | KF990036  | 1577 | 34 | 49  | 51                   |
| Pseudotsuga sinensis    | AB601120  | 1225 | 34 | 50  | 16                   |
| Psilotum nudum          | AP004638  | 1388 | 34 | 44  | 90                   |
| Pteridium aquilinum     | HM535629  | 1523 | 34 | 46  | 95                   |
| Ptilidium pulcherrimum  | HM222519  | 1190 | 33 | 50  | 36                   |
| Puelia olyriformis      | KC534844  | 1403 | 39 | 57  | 50                   |
| Quercus aliena          | KP301144  | 1699 | 37 | 51  | 40                   |
| Ramunculus macranthus   | DQ359689  | 1551 | 38 | 49  | 02                   |
| Raphanus sativus        | KJ716483  | 1533 | 36 | 47  | 90                   |
| Retrophyllum piresii    | KJ617081  | 1332 | 37 | 43  | 87                   |
| Rhazya stricta          | KJ123753  | 1547 | 36 | 49  | 56                   |
| Rosa odorata var. gigantea | KF753637 | 1566 | 34 | 49  | 51                   |
| Saccharum hybrid        | AP006714  | 1411 | 34 | 52  | 13                   |
| Salix interior          | KJ742926  | 1564 | 34 | 50  | 78                   |
| Salvia miltiorrhiza     | HF586694  | 1513 | 34 | 46  | 85                   |
| Sanonia uncinata        | KM111545  | 1243 | 29 | 43  | 25                   |
| Sarocalamus faberi      | JX513414  | 1396 | 39 | 57  | 13                   |
| Schefflera delavayi     | KC456166  | 1563 | 38 | 52  | 95                   |
| Schwabea americana     | HG738866  | 1699 | 38 | 51  | 23                   |
| Sedum sarmentosum       | JX427551  | 1504 | 38 | 47  | 96                   |
| Selaginella moellendorffii | HM173080 | 1437 | 51 | 50  | 84                   |
| Selaginella uncinata    | AB197035  | 1441 | 55 | 51  | 75                   |
| Sesamum indicum         | JN637766  | 1533 | 38 | 48  | 61                   |
| Setaria italica         | KJ001642  | 1388 | 39 | 56  | 32                   |
| Silene chalcedonica     | KF527886  | 1480 | 36 | 48  | 80                   |
| Sorghum bicolor         | EF115542  | 1407 | 38 | 57  | 71                   |
| Spiridela polyrhiza     | KJ160603  | 1687 | 36 | 48  | 80                   |
| Stangeria eriopus       | JX416858  | 1635 | 40 | 54  | 62                   |
| Stockwellia quadrijida  | KC180807  | 1595 | 37 | 50  | 36                   |
| Syntrichia ruralis      | FJ546412  | 1225 | 28 | 45  | 76                   |
| Taxus mairei            | KJ123824  | 1295 | 35 | 43  | 14                   |
| Tetracentron sinense    | KC608752  | 1644 | 38 | 49  | 29                   |
| Thamnocalamus spathiflorus | JX513425 | 1397 | 39 | 57  | 47                   |
| Trachelium caeruleum    | EU090187  | 1623 | 38 | 55  | 90                   |
| Trifolium subterraneum  | EU849875  | 1447 | 34 | 55  | 12                   |
| Trigonobalanus doichangensis | KF990556 | 1599 | 37 | 54  | 96                   |
| Trollium aethiopicum    | KJ912694  | 1149 | 37 | 51  | 40                   |
| Triticum aethiopicum    | KJ592713  | 1387 | 38 | 55  | 70                   |
| Triticum aethiopicum    | AB342240  | 1454 | 38 | 55  | 23                   |
| Trochodendron aralioides | KC609753  | 1659 | 38 | 50  | 46                   |
| Typha latifolia         | GU195652  | 1615 | 37 | 51  | 12                   |
| Utricularia gibba       | KC997777  | 1520 | 38 | 49  | 75                   |
| Vaccinium macrocarpon   | JQ248601  | 1760 | 37 | 72  | 42                   |
| Veratrum patulum        | KP437397  | 1539 | 38 | 49  | 23                   |
| Vigna radiata           | GC839327  | 1512 | 35 | 48  | 88                   |
| Vitis rotundifolia      | KF976463  | 1608 | 37 | 51  | 24                   |
| Vitis vinifera          | AD806289  | 1609 | 37 | 51  | 20                   |
| Vitis vinifera          | AD806290  | 1609 | 37 | 51  | 18                   |
| Vitis vinifera          | AD806291  | 1609 | 37 | 51  | 21                   |
| Vitis vinifera          | DQ424856  | 1609 | 37 | 50  | 93                   |
| Viviania marifolia      | KF240615  | 1572 | 38 | 59  | 61                   |
| Welwitschia mirabilis   | EU342371  | 1197 | 37 | 44  | 02                   |
| Wolliffia australiana   | JN100055  | 1687 | 36 | 43  | 94                   |
| Yushania levigata       | JX513426  | 1396 | 39 | 57  | 41                   |

continued on the next page
### A. Eight cluster structure of chloroplast genomes

Let now consider the patterns of chloroplast genomes. To do that, we just located the points corresponding to frequency dictionaries of the fragments of a chloroplast genome, in 63-dimensional space. Of course, there is no way to see this distribution immediately. We used ViDaExpert software [11] to visualize it. The best way to see a pattern provided by distribution of the fragments converted into frequency dictionaries is to see it in the space determined by three main principal components [29].

To begin with, we shall expand the labeling system described above. Previously, four labels have been introduced: phase 0, phase 1, phase 2 and junk. Now we add one more phase called tail, and split each phase $j^{th}$ into two subphases: these are the phases $F_0, F_1, F_2$, and $B_0, B_1, B_2$, respectively. The sense of these subphases is clear and apparent: they correspond to forward reading ($F_0, F_1$ and $F_2$) and backward reading ($B_0, B_1$ and $B_2$) of genes, in leader and ladder strands, respectively. The index coincides to the reminder of the division of the distance between the start position of a coding regions, and the center of a fragment, by 3.

The tail phase looks the most intriguing. First of all, it comprises the fragments falling into a dense series of tRNA (5S RNA, 25S RNA, etc.) genes. Probably, the points indicated as junk in the tail phase are the border fragments.

Now consider several genomes shown in principal components in two projection: in “full face” and in “profile”. The former means that the first principal component is normal to the plane of view, and the latter means that the first principal component is in the plane of view. Fig. 1 presents the “profile” view of the fragments distribution of ray grass ($Lolium perenne$, AC AM1777385 in EMBL-bank) genome, the coding regions. This is a typical “bullet-like” pattern of the distribution. In the right in this Figure the same distribution is shown with indication of the fragments labeled in color (see Figure legend). Junk phase is shown in the right figure in black.

Thus, eight clusters are distinctively identified, in this Figure: six ones correspond to six phases (from $F_0$ to $B_2$, respectively), the seventh cluster comprises the junk labeled fragments, and the eighth cluster (that is the tail) comprises the fragments of all seven mentioned above phases, while it is evidently distinguished from a main body of the distribution. This genome exhibits a typical “four-cluster” pattern, when seen in “full face”: the clusters corresponding to leader and ladder strand coincide, in the this projection mode.

Another very important feature of this genome is the clearly visible tail phase, in the distribution of fragments. This is rather frequent pattern observed among the studied genomes. The difference between the dictionaries $W^{(0)}_{(3,3)}$, $W^{(1)}_{(3,3)}$ and $W^{(2)}_{(3,3)}$ (see Table I) manifests in the clustering in “wings” (shown in color in Figs. 1 and 2); on the contrary, the lack of such difference observed for junk phase fragments results in a ball-shaped distribution of these points, in 63-dimensional space. Fig. 3 shows the junk phase fragments, only, with all the other erased from an image, in two projections. Unlike the bacterial genomes [11, 12], here junk exhibits the separation into two clusters (see Fig. 3(a)). It should be stressed that such two-cluster pattern of the distribution of the fragments falling into junk areas of a genome may not be observed, in general, without the fragments corresponding to coding areas of the genome. The question whether the junk phase fragments yield a pattern themselves, solely, is still open. Strictly speaking, this split of a junk fragments ensemble into a body of junk in proper sense, and in tail phase forces to claim the eight-cluster structure occurrence in chloroplasts genomes, in contrary to the patterns observed for bacterial genomes [11, 12].

Let now consider the fragments comprising the tail in more detail. To do that, we determine GC-content both for the entire genome, and for each fragment, and plot then the content against the number of a fragment. Fig. 4 shows this plot: the tail phase is colored in red. Let us remind, that the genome-wide GC-content of this entity is equal to 0.38. The overall GC-content has been reported to be the key factor defining the structure of clustering of the fragments formally identified within a bacterial genome [11, 12]; that former has significantly less effect, for chloroplast genomes. A tight examination of Table I shows that GC-content varies from 0.28 ($Orthotrichum rogeri$, AC KP119739 and Syntrichia ruralis, AC FJ544612) to 0.51 for Selaginella moellendorffii, AC HM173080 and even 0.55 for Selaginella uncinata, AC AB197035. Meanwhile, this Table says nothing about the specific values of GC-content of the fragments identified within various chloroplast genomes. Fig. 5 answer this question.

This Figure shows the set of chloroplast genomes under consideration (see Table I) ordered with respect to the genome-wide GC-content value. Beside, this Figure

| Gene                      | junk | tail |
|---------------------------|------|------|
| Genome                    | 0.9745 | 0.9617 | 0.6248 |
| Gene                      | 0.9218 | 0.6285 |
| junk                      | 0.5939 |
FIG. 1: The distribution of 12,244 fragments of *Lolium perenne* chloroplast genome (ACAM777385), “profile”, in principle coordinates. Left image presents a general overview, and the right one presents the phases $F_0$ – red triangles, $F_1$ – bright green diamonds, $F_2$ – light blue circles, $B_0$ – rosy triangles, $B_1$ – sand diamonds, $B_2$ – dark blue circles.

FIG. 2: Same genome as in Fig. 1 with junk phase erased; left is “profile” sight, and right is “full face” sight; color labeling is the same as in Fig. 1.

shows the plots of average GC-content determined over the ensemble of coding fragments (all six phases), of average GC-content of non-coding fragments, and of average GC-content of tail phase fragments. Evidently, the plots of genome-wide, coding and non-coding GC-content figures exhibit a high concordance in behaviour, while the tail phase fragments ensemble remains rather permanent.

Table III shows the correlations coefficients determined between averaged figures of GC-content of four ensembles of the fragments of genomes. The figures shown in Table III reveal the relative independence of the tail phase from the other parts of a genome, and GC-content of that latter never falls beyond 0.50 level. The set of genomes with lower figures of GC-content comprises the species *Orthotrichum rogeri*, *Syntrichia ruralis*, *Physcomitrella patens*, *Marchantia polymorpha*, *Sanionia uncinata*, *Anthoceros angustus*, *Ptilidium pulcherrimum*, *Equisetum arvense*, *Glycyrrhiza glabra*, *Trifolium sub-
FIG. 3: The same genome as in Fig. 1 junk phase shown in two projections, coding phases are erased.

FIG. 4: The plot of GC-content of all the fragments layered out alongside the *Lolium perenne* chloroplast genome; the tail phase (Fig. 3) is shown in red.


terraneum, Orobanche gracilis, Taxus mairei, Millettia pinnata, Pisum sativum, Juniperus virginiana and Juniperus bermudiana. The genomes of Anoaea mirabilis, Lygodium japonicum, Pteridium aquilinum, Ophioglossum californicum, Marsilea crenata and Myriopteris lindheimeri comprise the opposite group with higher figure of GC-content. Finally, two species (these are Selaginella moellendorffii and S. uncinata) yield the highest level of GC-content (see Table II for details).

Let now focus on the behaviour of the GC-content of the fragments comprising tails in the distribution of the fragments (see Fig. 5). Remarkably, there is no genome with GC-content figure lower than 0.5, for these fragments. Differing in this figure from the entire genome, the tails ensemble still comprises both coding, and non-coding fragments. The former are presented by densely located tRNA genes, 5S RNA and 16S RNA genes. This fact holds true for all genomes exhibiting a tail phase, and such genomes make a majority of entities studied in this paper.

Let now provide some examples of the fragments distributions observed in chloroplast genomes with various GC-content values. Consider the moss Physcomitrella patens (AC AP005672) genome with GC-content equal to 0.29 (next to the lowest one in the list of studied genomes). This is the model organism often used in evolutionary studies. Fig. 9 shows two projections of the full plot of the fragments; one easily can see similar pattern with two “tripods” overlapping each other, and the tail phase part. It should be stressed, that this genome exhibits another triplet with the least standard deviation figure: CGC, on the contrary to that one shown in Fig. 1. This genome exhibits stronger split of two phases (these are F1 vs. B0 and F0 vs. B1), in comparison to the pattern shown in Fig. 2 yet, the congruence of all the phases is strong enough.

To make it more clear, we show the distribution of all the fragments (the color labeling is the same as in Fig. 6) falling in coding regions, only; all the points corresponding to junk phase are erased. This figure allows to see that the third phase in this genome slightly deviates, in its clustering pattern, from two other couples: a reasonable part of blue and dark blue points belong to another cluster than that one comprising the third phase fragments. This distribution is shown in Fig. 10.

Fig. 7 shows the distribution of junk phase fragments of the moss genome. Similar to Fig. 5 this genome also exhibits an occurrence of some points of junk in tail phase. Whether it is a biologically sounding fact, still awaits for an answer. Again, it should be borne in mind, that all the distributions shown in Figs. 6 to 7 are not independent: actually, all these figures just show the same distribution, while some points are not shown in various figures; still, they affect the distribution pattern.

The patterns shown in Figs. 1 to 7 present a typical structuredness in a distribution of the small fragments of chloroplast genome. Actually, all the genomes except two entities exhibit such pattern in fragments distribution; these latter are the genomes of Selaginella moellendorffii (ACHM173080) and S. uncinata (AC AB197035). They are extremely ancient and rather isolated mosses belonging to primitive vascular plants. First of all, they have other triplets with the least standard deviation: GAC and TAA, respectively. Fig. 8 shows the distribution of all phases of S. moellendorffii genome. There is no tail phase at all, in this genome, neither in coding phases, nor in non-coding one. The pattern of distribution for S. uncinata is pretty close to that one shown in Fig. 8. Another indirect evidence for this issue is discussed in 10 (see also very useful paper 20).

B. Chloroplasts and cyanobacteria

The essential difference in the structuredness of a genome of chloroplast from bacterial genome is the key issue of the work. Still, the question arises whether this difference is essential. In other words, while chloroplasts form a tight and uniform group of genome bearers, bacteria are extremely diverse, both in genetics, phylogeny, taxonomy, physiology and ecology. What if there are some bacteria that had fallen out from our analysis, but still are close to chloroplasts, in the sense of the small fragments distribution? Indeed, the diversity of bacteria is huge, and there is no guarantee of the total absence of the coincidence of the structure described above when retrieved from some bacterial genome.

Speaking on the similitude or any other semblance of the patterns observed in chloroplast genomes to those observed in bacterial genomes, one should first of all concentrate on the comparison of the structures of chloroplasts, and cyanobacteria. These latter are stipulated to be the other branch of descendants of the common ancestor of chloroplasts and modern bacteria. Here we do not study this point in detail, while some preliminary results 33 show that the divergence between chloroplasts and cyanobacteria is tremendous. Fig. 9 illustrate the point.

IV. DISCUSSION

Let now get back to the labeling system (see page 3) of the formally identified fragments in a sequence. It provides a reasonable balance between the impact of coding and non-coding regions. Since the label value depends on the central nucleotide position, then approximately a half of the “border” fragments (i.e. those that cover the border between coding and non-coding regions in a genome) are labeled as junk, and another half are labeled as coding ones, with the specific phase value. Suppose, the total number of coding regions in a chloroplast genome is 50. Then an approximate number of “border” fragments labeled as junk is estimated as

\[
\frac{L}{2R} \times 50 \times 2 \approx 2500, \quad (5)
\]
FIG. 5: Average GC-content for entire genome, coding, non-coding parts and tail phase (cumulatively).

where the factor 2 counts both forward and backward oriented coding regions. The same number of the “border” fragments would be labeled with some phase figures; this parity guarantees, to some extent, a lack of distortion in the fragments clustering.

Papers [11, 12] present an approach to figure out a structuredness in bacterial genomes based on systemic and sequential comparison of frequency dictionaries of the fragments of a genome; the fragments were identified in the same way, as we have done. It should be stressed that such fragments were identified with neither respect to a functional charge of a fragment. The results presented in these papers show that the fragments tend to cluster located in the vertices of two triangles. The triangle vertices correspond to the phase of a fragment; in other words, a cluster comprises the fragments that have the same reading frame shift figure. A mutual arrangement of these two triangles is completely determined by the average (over the genome) figure of GC-content, for bacterial genomes.

A general seven-cluster structure was reported, for bacterial genomes, in these papers; the seventh cluster comprises the fragments falling into a junk area of the genome. The papers [11, 12] also provide an elegant explanation of an origin of this seven-cluster structuredness, describing the constraints forcing two triangles to rotate and project one over another. Here the genome-wide GC-content is claimed to be the only key factor determining the pattern of the cluster structure. A minor variation of GC-content results in visible change of the structure pattern.

There are following patterns of the fragments distribution, observed on bacterial genomes, for various figures of GC-content. GC-content close to 25% yields two “parallel triangles” (for AT-reach genomes); the growth of GC-content to ~ 35% yields the pattern with two “orthogonal triangles”, and the raise of GC-content up to 60% results in degeneration of two triangles into a single plane. Besides, the authors of [11, 12] claim such seven-cluster pattern be universal one; meanwhile, our results disprove this hypothesis.

A. Cluster structure of chloroplast genomes

Since chloroplasts take their origin from bacteria [15–18], then one may expect they inherit this universal pattern of the inner genome structuredness. Our re-
FIG. 6: The distribution of 11,118 fragments of moss *Physcomitrella patens* chloroplast genome (AC AP005672), left is “profile”, and right is “full face”: $F_0$ – red triangles, $F_1$ – bright green diamonds, $F_2$ – dark blue circles, $B_0$ – rosy triangles, $B_1$ – dark green diamonds, $B_2$ – blue circles.

FIG. 7: Junk-phase only distribution of moss chloroplast genome fragments.

Results disprove this assumption; moreover, GC-content of chloroplast genomes does not impact on the pattern of fragments distribution. The newly found pattern in small fragments distribution in 63-dimensional triplet frequency space seems to be very universal: there are two only exclusions from the list of studied genomes (see Table II). They are presented by two ancient moss species (*Selaginella moellendorffii* and *Selaginella uncinata*) originated more than $4 \times 10^8$ years ago.

The list of other genomes (see Table II) is apparently split into two parts: the former has the triplet GCG exhibiting the least standard deviation level, and the latter
has the triplet CGC. Indeed, these triplets exhibit a high similitude. The point is that they comprise so called complementary palindrome. Palindrome itself is a word that read equally in opposite directions (e.g., level in English; two words may form a palindromic couple (god ⇔ dog in English). The triplets GCG and CGC yield the so called complementary palindrome: a couple of two strings of DNA sequence that are read equally in opposite directions with respect to the Chargaff’s substitute rule (A ⇔ T and C ⇔ G). This really important symmetry in frequency dictionaries, but it completely falls beyond the scope of this paper.

Another important question here is whether the observed clusters corresponding to six phases (these are $F_0$, $F_1$, $F_2$, $B_0$, $B_1$ and $B_2$) really comprise clusters, or it is a kind of artifact resulted from a visualization technique. This question has obvious and transparent answer: yes, the clusters observed by visualization of the phases are the real clusters identified with a clustering technique. To check it, we have carried out $K$-means cluster implementation, of the frequency dictionaries corresponding to the fragments. Fig. 8 shows the clustering developed by $K$-means [29] (with $K = 4$) for the moss genome. Again, we did not aim to figure out some cluster structure due to $K$-means, but to verify the cluster structure observed in genomes through the visualization (that is the phase
Careful examination of Figs. 1, 2, 6 and 10 shows the general situation in localization of the phases, within a pattern. Indeed, the localization of the phases could be described by the following cyclic diagrams: $F_0 \rightarrow F_1 \rightarrow F_2 \rightarrow F_0$ (clockwise), and $B_0 \rightarrow B_1 \rightarrow B_2 \rightarrow B_0$ (counterclockwise). In fact, these two diagrams mirror each other, so that no complete coincidence make take place due to rotation. Such mirror symmetry corresponds to the double-stranded structure of DNA; the localization of $F_2$ and $B_2$ phases in the same projection is here of greater importance. All the studied chloroplast genomes exhibit such mirroring symmetry, while there are no evidences for that latter in bacterial ones [11, 12]. The phases $F_0$, $F_1$, $F_2$ make a triangle with given vertices circuit direction; same is true for the phases $B_0$, $B_1$ and $B_2$, and the circuit direction is the same, as for $F$-phases. This fact seems to be universal for bacteria (and some other genomes, e.g. fungi ones); on the contrary, chloroplast genomes exhibit exactly opposite pattern: they have counter-directed circuit directions, for those phases. Papers [37, 38] report on another type of structuredness found in chloroplast genomes, while we believe the mechanism staying behind these structures and those we are showing here, is the same: triplet frequency peculiarities. More specific mechanism based on codon bias yields a structuredness reported in [35]. These facts may reveal the “independent” evolution of chloro-
plast genomes (see also [34, 39]), on the contrary to the synchronized evolution of these latter with the host nuclear genome [19]. Also, such symmetry may answer the question towards the attribution of contigs for de novo assembling genomes [30–32] (see also another sight on the problem in [36]).

This mirroring has one more manifest in mutual location of the clusters comprising different phases. Fig. 12 illustrates this fact: while the location of phase 0 and phase 1 remains the same, in both subfigures, the location of the phase 2 takes mirroring positions. The phase 2 cluster faces down, for Anthoceros angustus, and

FIG. 11: $K$-means ($K = 4$) for Physcomitrella patens (ACAP005672) chloroplast genome.

FIG. 12: “Up” and “down” orientation of the clusters shown on two genomes: Ranunculus macranthus (up), and Anthoceros angustus (down).
TABLE IV: Distribution of patterns.

| Pattern | frequency |
|---------|-----------|
| GCG     | 19        |
| GCG     | 20        |

that former faces up for large buttercup (*Ranunculus macranthus*). To make the images more apparent, we have erased the points corresponding to junk. Two positions of phase 2 cluster correspond to two mirroring axes systems. Let now get back to Table II the last column in the table (labeled with ▼ sign) indicates the orientation of the clusters: U stands for “up” positioning of the cluster, and D stands for “down” positioning of that latter. Comparing Figs. 2, 10, 6, 12 to Fig. 9 (see Subsec. III B), one sees that such mirroring symmetry is universal, for chloroplast genomes: cyanobacteria that are claimed to be evolutionary related to chloroplasts, do not exhibit such pattern, at all.

Another sounding manifestation of the symmetry is the interchange of the triplet yielding the least standard deviation figure; see again Table II Indeed, with exclusion of two of triplets (these are GAC and TAA), all other entries exhibit either GCG, or CGC triplet yielding the least standard deviation figure. The species with unconventional triplets GAC and TAA are actually the ancient moss organisms of *Selaginella* genus appeared app. 400 billion years ago. The unconventionality of the triplets yielding the least standard deviation figure may result from this long isolated lineage.

Apart these two species, all other ones (see Table II) could be split into two groups: the former with GCG triplet yielding the least standard deviation, and the latter with CGC triplet; the abundances of each groups are 115 and 61 entries, respectively. It should be mentioned that two genomes were not annotated, completely; thus, we were not able to determine what type of symmetry they exhibit. Table IV summarizes the distribution of chloroplast genomes over the combinations of U ↔ D variants, and the triplets CGC ↔ GCG. In such capacity, the genomes with CGC triplet differ from those with GCG ones. Whether this difference is of a nature of things, or results from a bias of the database used in the study, should be examined further. One definitely could say there is no correlation between the pattern of orientation, triplet with the least standard deviation figure, and separation of plants on gymnosperm vs. angiosperm species (cf. Table IV and Table II).

### B. Specific type of symmetry and coding regions

Consider now the abundances of the beams (or clusters) corresponding to the phases $F_0$ through $B_2$; obviously, the must be equal, or at least pretty close, since the beams differ in the reading frame shift of a triplet, only. Typical figures are the following: $|F_0 - B_0| = 2488, 4$, $|F_1 + B_1| = 2488, 6$, and $|F_2 + B_2| = 2485, 8$. The greatest standard deviation of the beam abundances is provided by *Hibiscus syriacus* (AC KP688069 in EMBL–bank), and the figure is 14.53. Reciprocally, the least figure (that is exactly zero) is provided by *Olimarabidopsis pumila* (AC AP009368).

The difference between the phases $|F_0 - B_0|$, $|F_1 - B_1|$ and $|F_2 - B_2|$ are of greater interest. These figures vary from $-1305$ (averaged over three beams), for *Ophioglossum californicum* (AC KC117178) to 1387, for *Equisetum arvense* (AC GU191334). Fig. 13 shows the relation of the bias in forward and backward coding regions occurrence, in different organisms, and the type of their mirroring symmetry. This figure shows the set of genomes (Table IV) ordered ascending on $|F_0 - B_0|$ figures; in other words, the left genome has $|F_0| - |B_0| = -1362$ (that is *Ophioglossum californicum*, AC KC117178), while the right one exhibits $|F_0| - |B_0| = 1382$ (that is *Equisetum arvense*, AC GU191334). The solid black line in this Fig. shows the standard deviation of cluster abundances determined over all six phases; small red diamonds show the symmetry orientation: upper dots show U type, and lower ones show D type. It seems that the excess of the abundance of the fragments belonging to backward phases over those belonging to forward phases in 600 entities results in the unambiguous determination of U type symmetry orientation; the right end of this curve supports, to some extent, this idea.

### C. What is beyond?

The study of statistical properties of DNA sequences still challenges researchers, and may bring a lot new. Here we have demonstrated basic structural difference of chloroplast genomes from the bacterial ones, manifested in the clustering in distribution of small formally identified fragments of a genome. Below are some issues that had fallen from the scope of this paper, while they are rather important to be considered in the nearest future.

#### 1. Dark matter of a genome

Functional and evolution roles of the junk in a genome still is conspired from researchers. It is extremely doubtful that junk has no matter in a genome, neither it plays no role in various and complicated biological processes involved into an inherited information processing and functioning. For some cases (see [30], the removal of junk enforces the clustering of coding regions and makes easier the comprehension of the peculiarities standing behind. Yet, special efforts must be addressed to reveal the role and impact of junk regions of a genome on the processes mentioned above.

A variety of aspect of the influence of a junk on clustering observed within a genome is very wide. Not speaking about the differences in statistical properties of frequency dictionaries $W_{(3,3)}$ (and $W_{(m,n)}$, in general) ob-
served for junk fragments of a genome vs. those observed for coding ones, one may expect even the strong impact from the ratio of coding/non-coding parts occurred within a genome. For instance, here we report on mirror symmetry in mutual interlocation of six coding phases, for the frequency dictionary $W_{(3,3)}$ developed for chloroplast genomes. Fig. 9 explicitly demonstrates an absence of such symmetry, for cyanobacteria genome, and this fact may result from a significant difference in the coding/non-coding ratio figures observed for these genetic systems.

2. Other chloroplast genomes

Here we present some results obtained on the careful examination of 178 genomes of ground plants. Yet, the generality of the observation awaits for further approval: first of all, one should study the chloroplast genomes of the organisms deviating rather far from the studied ones, in ecology (water plants, and algae, especially), physiology, taxonomy. Such systemic examination is the matter of the nearest future work of ours.

V. CONCLUSION

Here we studied the structuredness of chloroplast genomes revealed through the clustering of frequency dictionaries of considerably short fragments of a genome that were determined formally, with neither respect to the function encoded in a part of the genome fell into the fragment. The triplet dictionaries were developed, to cluster; these former counts triplets with no overlapping, while with no gaps between any two triplets. The fragments are distributed into eight distinct clusters: six of them gather the fragments falling into the coding regions, and differ in reading frame shift; the shift manifests in phase index of a fragment. The seventh cluster comprises the fragments falling into non-coding regions, and finally, he eighth cluster (so called tail) comprises the fragments with excessive GC-content value. These fragments correspond to the region where various tRNA and SRNA genes are concentrated; probably, this cluster includes also the “border” fragments (those that contain a border between coding and non-coding parts of a genome).

The clusters exhibit wonderful mirroring symmetry:
the phase circuit in the forward and backward strands are counter-directed; this fact completely contradicts to the similar structure observed for bacteria, including cyanobacteria (which are stipulated to be the descendants of a common ancestor with chloroplasts). Such mirror symmetry yields a separation of the genomes into two groups: those with “up”-directed location of the cluster comprising $F_2$ and $B_2$ phases vs. those with “down”-directed; apparently, the threshold in the abundances of the phases gathered into a single cluster determines the direction of the $F_2 \div B_2$ cluster.

Acknowledgments

We would like to extend our gratitude to Prof. Alexander N. Gorban from Leicester University and Andrew Yu. Zinovyev from Curie Institute for long-time collaboration and permanent encouraging interest to the work. The work was partly supported by the grant from Russian Government (# 14.Y26.31.0004).

[1] Dong W., Liu J., Yu J., Wang L., Zhou Sh. Highly Variable Chloroplast Markers for Evaluating Plant Phylogeny at Low Taxonomic Levels and for DNA Barcoding PLoS ONE. 2012 7(4):1–9.
[2] Gielly L., Taberlet P. The Use of Chloroplast DNA to Resolve Plant Phylogenies: Noncoding versus rbcL Sequences Molecular Biology and Evolution 1994 11(5):769–77.
[3] Chun J., Rainey F.A. Integrating genomics into the taxonomy and systematics of the Bacteria and Archaea Int. J. of Systematic and Evolutionary Microbiol. 2014 64:316–24.
[4] Franklin L.R. Bacteria, Sex, and Systematics Philosophy of Science 2007 74:69–95.
[5] Carbonell-Caballero J., Alonso R., Ibanez V., Terol J., Tulon M., Dopazo J. A phylogenetic analysis of 34 chloroplast genomes elucidates the relationships between wild and domestic species within the genus citrus Mol. Biol. Evol. 2015 32(8):2015–35.
[6] Leliært F., Smith D.R., Moreau H., Herron M.D., Verbruggen H., Delwiche Ch.F., De Clerck O. Phylogeny and Molecular Evolution of the Green Algae Critical Reviews in Plant Sciences 2012 31:1–46.
[7] Milanowski R., Zakrys B., Kwiatkowski J. Phylogenetic analysis of chloroplast small subunit rRNA genes of the genus Euglena Ehrenberg International Journal of Systematic and Evolutionary Microbiology 2001 51:773–81.
[8] Katayama H., Oghara Y. Phylogenetic affinities of the grasses to other monocots revealed by molecular analysis of chloroplast DNA Current Genetics 1996 29:572–81.
[9] Marazzi B., Endress P.K., De Queiroz L.P., Conti E. Phylogenetic relationships within senna (leguminosae, cassinae) based on three chloroplast DNA regions: patterns in the evolution of floral symmetry and extrafloral nectaries American Journal of Botany. 2006 93(2):288–303.
[10] Shaw J., Lickey E.B., Beck J.T., Farmer S.B., Liu W., Miller J., et al. The tortoise and the hare ii: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis American Journal of Botany 2005 92(1):142–66.
[11] Gorban A.N., Zinovyev A.Yu., Popova T.G. Seven clusters in genomic triplet distributions Silico Biology. 2003 3(4):471–82.
[12] Gorban A.N., Zinovyev A.Yu., Popova T.G. Four basic symmetry types in the universal 7-cluster structure of microbial genomic sequences Silico Biology. 2005 5(3):265–82.
[13] Cavalier-Smith T. Chloroplast Evolution: Secondary Dispatch Symbiogenesis and Multiple Losses. Curr. Biol. 2002 12:62–64.
[14] Falcón L.I., Magallón S., Castillo A. Dating the cyanobacterial ancestor of the chloroplast. ISME J. 2010 4:777–83.
[15] Mereschkowsky K.S. Theorie der zwei Plasmaarten als Grundlage der Symbiogenesis, einer neuen Lehre von der Entstehung der Organismen. Biol Centralbl. 1910 30:353–67.
[16] Mereschkowsky K.S. ¨Uber Natur und Ursprung der Chromatophoren im Pfl anzen reiche Biol. Zentr.-Bl. Bd. 1905 85(18):593–604.
[17] Zimorski V., Ku Ch., Martin W.F., Gould S.B. Endosymbiotic theory for organelle origins Current Opinion in Microbiology 2014 22:38–48.
[18] Raven J.A., Allen J.F. Genomics and chloroplast evolution: what did cyanobacteria do for plants? Genome Biology 2003 4(3):209.
[19] Sadovsky M.G., Putintseva Yu.A., Chernysheva A.I., Fedotova V.S. Genome Structure of Organelles Strongly Relates to Taxonomy of Bearers LNBI 2015 9043:482–90.
[20] Sato N. Comparative Analysis of the Genomes of Cyanobacteria and Plants Genome Informatics 2002 13:173–82.
[21] Fahad A., Alshatri N., Tari Z., Alamri A., Khalil I., Zomaya A.Y., et al. A survey of clustering algorithms for big data: taxonomy and empirical analysis IEEE Trans. on emerging topics in computing. 2014 2(3):267–279.
[22] Dongkuan Xu, Yingjie Tian A Comprehensive Survey of Clustering Algorithms Ann. Data. Sci. 2015 2(2):165–193.
[23] Gorban A.N., Kegl B., Wünsch D.III, Zinovyev A.Yu. (Eds.) Principal Manifolds for Data Visualisation and Dimension Reduction LNCSE 2007 58, Springer, Berlin – Heidelberg – New York.
[24] Gorban A.N., Roose D. (Eds.) Coping with Complexity: Model Reduction and Data Analysis Lecture Notes in Computational Science and Engineering 2010 Springer.
[25] Gorban A.N., Zinovyev A.Yu., Wünsch D.III Application of the method of elastic maps in analysis of genetic texts Proceedings of International Joint Conference on Neural Networks, 2003 3:1826–31.
[26] Gorban A.N., Zinovyev A.Yu. Principal Graphs and Manifolds Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods and Techniques, Olivas E.S. et al Eds. Information Science Reference, IGI Global: Hershey, PA, USA 2009 28–59.
[27] Gorban A.N., Zinovyev A.Yu. Principal manifolds and graphs in practice: from molecular biology to dynamical systems Int. J. of Neural Systems 2010 20(3):219–32.
[28] Gorban A.N., Zinovyev A.Yu. Elastic principal manifolds and their practical applications Computing 2005 75(4):359–79.
[29] Fukunaga K. Introduction to statistical pattern recognition. Academic Press 1990. 591 p.
[30] Sadovsky M.G., Putinseva Yu.A., Birukov V.V., Krutovsky K.V. De novo assembly and cluster analysis of Siberian larch transcriptome and genome LNBI 2016 9656:455–64.
[31] Sadovsky M.G., Putinseva Yu.A., Krutovsky K.V., Orekhova N.V., Bondar Eu.I., Vaganov E.A. Seven-Cluster Structure of Larch Chloroplast Genome SFU Journal, ser. Biology 2015 8(3):268–77
[32] Sadovsky M.G., Putinseva Yu.A., Krutovsky K.V., Orekhova N.V., Birukov V.V., Vaganov E.A. Symmetry of Siberian Larch Transcriptome SFU Journal, ser. Biology 2015 8(3):27–86
[33] Sadovsky M.G., Senashova M.Yu., Malyshev A.V. Eight cluster structuredness of genomes of ground plants Russian J. of Gen.Biol. 2018, 79(2): in press
[34] Kaila T., Chaduvla P.K., Rawal H.C., Saxena S., Tyagi A., Mithra S.V.A. et al. Chloroplast Genome Sequence of Clusterbean (Cyamopsis tetragonoloba L.): Genome Structure and Comparative Analysis Genes 2017, 8, 212
[35] Mazumdar P., Othman R.Y.B., Mebus K., Ramakrishnan N., Harikrishnan J.A. Codon usage and codon pair patterns in non-grass monocot genomes Annals of Botany 2017 120(6):893-909
[36] Lehtonen S., Mylly L., Huttunen S. Phylogenetic analysis of non-coding plastid DNA in the presence of short inversions Phytotaxa 2009, 1: 3-20
[37] Morton B.R. Strand asymmetry and codon usage bias in the chloroplast genome of Euglena gracilis Proc. Natl. Acad. Sci. USA 1999, 96:5123-5128
[38] Bélauger A.-S., Brouard J.-S., Charlebois P., Otis C., Lemieux C., Turmel M. Distinctive architecture of the chloroplast genome in the chlorophycean green alga Stigeoclonium helveticum Mol Gen Genomics (2006) 276:464–477
[39] Cavalier-Smith T. Chloroplast Evolution: Secondary Symbiogenesis and Multiple Losses Current Biology, 2002, 12:R62-R64
[40] Falcón L.I., Magallón S., Castillo A. Dating the cyanobacterial ancestor of the chloroplast The ISME Journal, 2010, 4:777-783
[41] http://bioinfo.curie.fr/projects/ViDaExpert