 doi:10.20944/preprints202201.0425.v1

Article

Function Overcomes Taxonomy for Organella Genes Triplet Composition

Michael Sadovsky[1,2,4]*, Maria Senashova[1], Olga Mutovina[3], Anna Molyavko[1,3], Viktoria Fedotovskaya[3], Tatiana Shpagina[3], Yana Nedorez[3] and Yulia Putintseva[3]

[1] Institute of computational modelling SB RAS; 660036 Russia, Krasnoyarsk, Akademgorodok, 44, bldg. 55 *{msad,msen,annamo}@icm.krasn.ru*
[2] V.F. Voino-Yasenetsky Krasnoyarsk state medical university; 660022 Russia, Krasnoyarsk, Partizana Zheleznjaka str., 1
[3] Siberian federal university; 660041 Russia, Krasnoyarsk, Svobodny prosp., 79 *mutovina.ole4ka@mail.ru, viktoriia.fedotovskaia@gmail.com, shpaginat98@gmail.com, nedorez.ya@gmail.com, yuliya-putintseva@rambler.ru*
[4] Federal Research & Clinic Center of FMBA of Russia 660037 Russia, Krasnoyarsk, Kolomenskaya str., 26
* Correspondence: msad@icm.krasn.ru; Cell tel.: +7-902-990-4597 (M.S.)

Abstract: A comprehensive presentation of a variety of biologically sounding properties of genomes is present; chloroplast genomes are used as a biological matter. Triplet frequency composition is the general issue standing behind the properties. Besides, the new alignment-free error tolerant method of sequences comparison highly efficient for in/del mismatches is present, for transposons search. Triplet frequency dictionaries determined for a genome, or for a part of that latter were studied through various clustering techniques. The interplay between triplet composition and function reveals on tRNA genes unambiguously shows the prevalence of the function encoded in tRNA gene over the phylogeny: the genes gather into the clusters comprising the genes encoding the same amino acid; more exactly, few gene families exhibit fine cluster pattern corresponding the synonymous codons of an amino acid. Previously reported symmetry in chloroplast genomes is shown for a set of gymnosperm: that is mirror symmetry, rotational symmetry and the second Chargaff's parity rule asymmetry. A family of transposons was found in gymnosperm chloroplast genomes. This family is revealed through the novel comparison method based on convolution calculation, for a set of DNA sequences.

Keywords: transposon; order; triplet frequency; tRNA; clustering; taxonomy; symmetry; photosystem

1. Introduction

An exploration of the interplay between the structure of genomic entities, the function encoded in them, and the taxonomy of their bearers is the crucial issue in up-to-date molecular biology, molecular genetics, and bioinformatics. A genome is a complex object of immense size; however, plants have a specific part of a genome that belongs to chloroplasts.

Identity of the function encoded in chloroplasts is of great value for studying the interplay mentioned above: one meets no effects resulting from the difference in function encoded in the genome. That is true for all organelle genomes. Thus, a three-side problem is reduced to the two-side one: there is an interplay between structure (whatever it could be) and taxonomy of bearers of the chloroplast genomes. Comprehensive analysis of the interplay between structure and taxonomy in chloroplasts may reveal the details of the evolution of various species of plants.

Speaking on evolution, one must keep in mind the relative independence of the genetic entities: the chloroplast genome is separated physically from the nuclear one. Of course, the interaction of these two genomes is significantly more potent than that observed between the nuclear genome and mitochondrial one.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2022 doi:10.20944/preprints202201.0425.v1

1.1. Transfer RNA (tRNA) of chloroplast genomes: between function and taxonomy

Transfer RNAs (tRNA) are the moleculae playing the pivotal role in the biosynthesis of peptides: they carry on amino acid residues to ribosomes where the synthesis runs. These RNAs are encoded within a genome; chloroplast genomes have specific transfer RNAs. tRNA genes are relatively short (typical length is about 70 b. p.). We shall refer *t-genome* to a set of all tRNA genes of an organelle. A consistent and comprehensive study of a t-genome may bring a lot to understanding some evolution processes since a t-genome seems to be evolutionary very stable.

Also, a t-genome makes a highly suitable ensemble of genetic entities to reveal the relation between structure and function; here, triplet frequency dictionary (see subsec. 2.1) is stipulated to be a structure. The function of tRNAs is apparent: to transfer various amino acid residues to ribosomes. Luckily, a function could be considered in two aspects: the former is a specific amino acid correspondence, and the latter is synonymity of the codons encoding the same amino acid. This two-level pattern in function may enlighten some conspired details in structure–function interplay.

So, the question is as follows. Given a set of t-genomes of chloroplasts of several species, find whether tRNAs responsible for transferring a specific amino acid residue tend to exhibit proximal structure patterns or not; the same question arises to specific codons from a family of synonyms. The proximity of structure patterns is revealed by clustering triplet frequency dictionaries derived from the relevant genes.

Speaking in advance, function (that is, a specific amino acid residue transfer) beats taxonomy: being converted into triplet frequency dictionary, tRNAs tend to gather into relatively dense clusters so that a cluster comprises the genes of the tRNA transferring the same amino acid. Moreover, some tRNA with synonym codons exhibit an arrangement into a single cluster.

1.2. Chloroplast photosystem genes: between function and taxonomy

Sun is the primary source of energy on Earth, and many organisms are adapted to use it for their needs. Plants, algae, and cyanobacteria grow up due to the ability to consume sunlight; this is how photosynthesis and, accordingly, phototrophic nutrition runs. Photosynthesis genes seem to be relatively stable and conservative; therefore, they perfectly meet the study of the interplay between structure (triplet composition), function (various proteins to be encoded), and taxonomy of the bearers of genes.

We studied Photosystem I and Photosystem II (PS I and PS II) genes. They are located in chloroplast membranes (PS II is located in compressed granular membranes and PS I is located in uncompressed stromal membranes) [2, 12]. We investigated the distribution of chloroplast photosystem genes belonging to PS I and PS II in the space of triplet frequencies (see Subsec. 2.1 for more details). A nucleotide sequence conversion into triplet frequency dictionary transforms it into a point in 63-dimensional Euclidean space, so the distribution of these points in the space yields a structure. However, one may see many other definitions of a structure. Further, we focus on the structure provided by triplet frequency dictionaries only. This approach is close to that one presented in Subsec. 1.1 and provides the patterns of the interplay between structure, function, and taxonomy observed over another genetic material.

1.3. ATP synthase genes and $NADH^+$ genes of chloroplasts

Also, we applied the same methodology to study the interplay of structure (triplet dictionaries), function, and taxonomy of the bearers of two families of genes: the former is ATP synthase genes family, and the latter is $NADH^+$ genes family retrieved from chloroplast genomes of various species.

The key idea of these studies is to verify the hypothesis towards the prevalence of a function over taxonomy for various gene systems. We converted gene sequences into the triplet frequency dictionaries, then studied the distribution of the corresponding points in 63-dimensional Euclidean space to check whether the points were arranged

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2022 doi:10.20944/preprints202201.0425.v1

into clusters (apparently identified separated groups). If clustering takes place, then the composition of the clusters was studied: there might be two options. The former is that a cluster preferably comprises the genes of various functions but belonging to the same species. The latter is that a cluster preferably comprises the genes of the same function but belonging to various species. The first option makes taxonomy the leading factor determining the cluster composition, and the second option makes the function this factor. Speaking in advance, the substantial prevalence of the function over taxonomy has been observed.

1.4. Transposons in chloroplast genomes

Transposons [1–3] are the (moderately short) subsequences occurred in DNA playing essential role in a number of processes of intracellular regulation [4], evolution and genetic information processing [5–8]. They have been found in a great number of genomes of species ranging in taxonomy from bacteria [7] to man [9,10]. However, chloroplasts are known for their quite distinguished behaviour in transposons occurrence [3,4,7,11–14]. Probably, for the sake of exactness, one should say that a typical pattern of TE occurrence in chloroplast genomes differs from similar ones observed in other genetic entities [14].

However, transposons are found in chloroplast genomes. In such a capacity, one should examine the difference between transposons found in chloroplasts and those found in other genetic entities. Here we pursue this idea through the implementation of two independent methods of subsequence search in DNA; the former is based on classical alignment, and the latter is based on convolution calculation for specially pretreated DNA sequences (see subseq. 2.4 for the details of that method and subsec. 3.5 for the results).

Transposons are known for their ability to "jump" in a genome and to increase in a number of copies; their effect is not always positive in terms of a species survival [15]. The crucial role of transposons is probably in affecting the genes expression located around the transposon-containing site. They promote the chromosomal rearrangements due to recombinations and change methylation patterns through epigenetic pathways [15].

The novel method has no parameters to be adjusted. It is a very significant advantage since it provides a highly standardized comparison free from any arbitrariness of a parameter choice or involuntary implementation of some extra knowledge or constraints into the comparison procedure.

1.5. Intergenic subsequences in chloroplast genomes

We shall refer the genome subsequences encoding other products than a protein to intergenic DNA. An abundance of that former varies significantly both in various taxa and within a set of organisms of the same species [16]. The role and meaning of intergenic fragments of DNA are still arguable. Paper [17] promotes a theory saying that such fragments are a kind of evolutionary "buffer" which supports the stability of a genome. Some entities are very stable, thus tracing their presence back into a billion generations. Also, there is evidence that intergenic fragments may determine some diseases [18].

Silent DNA is widely used in phylogeny since it is stipulated to be similar in different species. Pseudogenes may yield new genes since they are expected to be less conservative if compared to coding DNA so that mutations take place in them preferably [19].

Some researchers claim the intergenic DNA fragments are less affected by selection; anyway, it is a common fact that their triplet structures differ significantly from that one observed for genes. These former exhibits a large-scale correlation in nucleotide location, while the coding DNA exhibit short-range correlations. Also, intergenic fragments are hypothesized to vary significantly within a genus, in dependence on the growth conditions [20,21].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2022 doi:10.20944/preprints202201.0425.v1

2. Materials and Methods

2.1. Frequency dictionary

Structures found in DNA moleculae are very diverse; there is no way to list all of them here. Everywhere further (except the section 1.4) we focus on the specific structural entity called *frequency dictionary* $W_{(q,t)}$. It is the list of all q-tipples counted along a sequence, if reading window identifying a string moves forward with the step t. We shall keep ourselves with triplet frequency dictionaries $W_{(3,1)}$ and $W_{(3,3)}$ in our study (see [22–24] for details).

Triplet frequency dictionary $W_{(3,1)}$ (or W_3, for the sake of brevity) is the list of all 64 triplets accompanied with the frequency of each entry. To do it, the numbers n_ω of each triplets $\omega = \nu_1\nu_2\nu_3$ are counted along a sequence under consideration; then the numbers n_ω are changed for the frequencies

$$f_\omega = \frac{n_\omega}{M}\,, \tag{1}$$

where M is the total number of all triplets met in a sequence. If a sequence is connected into a ring, M coincides with the length of the sequence. The triplets in a dictionary $W_{(3,3)}$ are counted with the step 3; in other words, there are neither gaps nor overlaps in the reading frame positions identifying the triplets here. This dictionary coincides to codon count if determined over a coding region (say, mature RNA); however, it should be borne in mind that we would count this type of dictionary for formally defined fragments of a sequence, and in such capacity, it may not be identical to codon count.

Any triplet frequency dictionary maps a sequence into a point in 63-dimensional space. Indeed, 64 triplets meet the linear constraint

$$\sum_{\omega=\mathsf{AAA}}^{\mathsf{TTT}} f_\omega = 1\,. \tag{2}$$

The constraint (2) allows 63 (linearly independent) triplets, thus making the space 63-dimensional.

Generally, a frequency dictionary $W_{(q,t)}$ is defined in the following way. For a given DNA sequence, put a window of the length q at the first nucleotide in the sequence and fix the identified string (word) into the list of entries. Then move the window along the sequence for t nucleotides, and fix the next word into the list. Go on in this way while the last complete coverage is found. The total number of words in a dictionary is then $\approx \frac{N}{t}$, where N is a sequence length. Counting the number n_ω of identical words ω, one gets a finite dictionary; changing numbers for frequencies

$$f_\omega = \frac{n_\omega}{\Omega}, \quad \Omega = \sum_\omega n_\omega\,, \tag{3}$$

one gets the frequency dictionary $W_{(q,t)}$. Obviously, there exists t (different, in general case) dictionaries $W^j_{(q,t)}$, where $0 \le j \le t-1$ determined for t different starting positions of the window of the length q; practically, we used $t = 1$. A conversion of a genetic sequence into a triplet frequency dictionary (regardless of the step t value) maps the sequence into a metric space; in such capacity, metrics must be introduced. We used conventional Euclidean metrics $\rho(\mathcal{S}_i, \mathcal{S}_j)$

$$\rho(\mathcal{S}_i, \mathcal{S}_j) = \sqrt{\sum_{\omega=\mathsf{AAA}}^{\mathsf{TTT}} \left(f_\omega^{(i)} - f_\omega^{(j)}\right)^2} \tag{4}$$

in our studies; here $f_\omega^{(i)}$ ($f_\omega^{(j)}$, correspondingly) is the frequency of a triplet ω counted over $\mathcal{S}_i$ (over $\mathcal{S}_j$, correspondingly).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2022 doi:10.20944/preprints202201.0425.v1

2.2. Genetic material

We used both GenBank and EMBL–bank to retrieve the genomes. All manipulations with this latter (gene extraction, etc.) have been done due to annotation.

2.2.1. T-genome description

We studied the sets of tRNA genes of chloroplast genomes of gymnosperm plants. The database enlists 4015 genes, totally; Table 1 shows the abundances of the genes encoding various amino acids. Besides, the table shows the coloring scheme (RGB, to be exact) used elsewhere in the Figures. This Table provides the numbers of genes encoding specific amino acids regardless of the synonymity of these latter. Also, this Table shows the labels used in the section on t-genome analysis for identifying amino acids in Figures.

Table 1. Genes abundance and amino acids coloring scheme. aa is amino acid, N is abundance of genes, RGB is coloring scheme and $*$ stands for symbol used as map label.

aa	N	RGB	$*$	aa	N	RGB	$*$	aa	N	RGB	$*$
A	129	255, 255, 0	△	I	301	153, 50, 204	○	Q	166	127, 255, 212	◇
C	145	255, 165, 0	□	K	141	147, 112, 219	△	R	371	50, 205, 50	○
D	148	255, 0, 0	◇	L	419	30, 144, 255	□	S	427	0, 255, 0	△
E	143	230, 230, 250	○	M	146	0, 0, 255	◇	T	302	0, 128, 0	□
F	143	199, 21, 133	△	fM	142	135,206,250	○	V	271	222, 184, 135	◇
G	258	255, 105, 180	□	N	144	0, 206, 209	△	W	148	128, 128, 128	○
H	158	255, 0, 255	◇	P	275	0, 255, 255	□	Y	148	192, 192, 192	△

Table 1 shows a reasonable abundance of genes set. However, the distribution of the genes over the set of the synonyms specific for each amino acid is extremely biased. It should be stressed, that some isodecoders are absent at the ensemble of genes; following tRNA genes with the anticodons are absent: **A** – CGC, **G** – ACC, **I** – UAU, **K** – CUU, **L** – AAG, **L** – CAG, **L** – GAG, **N** – AUU, **P** – AGG, **Q** – CUG, **R** – GCG, **R** – UCG, **S** – ACU, **T** – AGU; this discrepancy is discussed above.

Table 2. Number of genes of various amino acids, with respect to synonyms.

aa	ω	N	aa	ω	N	aa	ω	N	aa	ω	N	aa	ω	N
A	UGC	98	G	GCC	87	L	UAA	98	R	UCU	115	V	GAC	109
C	GCA	124	H	GUG	126	M	CAU	112	R	CCG	55	V	UAC	105
D	GUC	103	I	CAU	123	N	GUU	114	S	GCU	118	W	CCA	116
E	UUC	113	I	GAU	107	P	UGG	113	S	UGA	113	Y	GUA	115
F	GAA	114	K	UUU	112	P	GGG	102	S	GGA	99			
fM	CAU	110	L	UAG	119	Q	UUG	138	T	GGU	132			
G	UCC	114	L	CAA	113	R	ACG	121	T	UGU	106			

Table 2 shows the number of genes encoding specific amino acid; aa stands for amino acid, ω is the triplet, and N is the number of the genes encoding the specific amino acid. Colorless cells correspond to the genes where a single isodecoder is present; in the case of synonymity, the synonyms are colored in red and green (for the case of two synonyms), red, green, and blue (for the case of three synonyms). The synonyms, if any, are ordered descendingly with respect to the number of the genes.

2.2.2. Photosystem genes

The genes of photosynthetic system I and II were isolated from 570 chloroplast genomes currently available in the EMBL–bank. The following genes were found in the studied set of genomes: *psa*A, *psa*B, *psa*C, *psa*I, *psa*J, *psa*M, *psb*A, *psb*B, *psb*C, *psb*D, *psb*E, *psb*F, *psb*G, *psb*H, *psb*I, *psb*J, *psb*K, *psb*L, *psb*M, *psb*N, *psb*T. Here *pca* stands for PS I and *psb* stands for PS II. We have studied 474 chloroplast genomes, totally, in this section.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2022 doi:10.20944/preprints202201.0425.v1

2.2.3. ATP synthase genes and $NADH^+$ genes

Along the photosystem genes, we studied the distribution of two types of genes from chloroplast genomes: these are ATP synthase genes and $NADH^+$ genes. 84 species were analyzed, with the following species distribution in families: *Fabaceae* comprises 19 species, *Malvaceae* comprises 12 species, *Brassicaceae* comprises 8 species, *Poaceae* and *Salicaceae* both comprise 4 species. The families of *Anacardiaceae*, *Arecaceae*, *Asteraceae*, *Cucurbitaceae*, *Euphorbiaceae* and *Solanaceae* comprise 2 species, each. Finally, *Apiaceae*, *Apocynaceae*, *Araceae*, *Campanulaceae*, *Caricaceae*, *Caryophyllaceae*, *Cycadaceae*, *Gesneriaceae*, *Ginkgoaceae*, *Lamiaceae*, *Lentibulariaceae*, *Lythraceae*, *Magnoliaceae*, *Moraceae*, *Myrtaceae*, *Nelumbonaceae*, *Nyctaginaceae*, *Passifloraceae*, *Rhamnaceae*, *Rosaceae*, *Rutaceae*, *Sapindaceae*, *Theaceae*, *Vitaceae* and *Welwitschiaceae* families are presented with a single species.

The ATP synthase, and $NADH^+$ genes were retrieved from whole genomes deposited in NCBI bank using CLC workbench and converted to $W_{(3,1)}$ frequency dictionaries with *ad hoc* software. The databases comprise six types of ATP synthase genes, and 11 types of $NADH^+$ genes, respectively.

2.2.4. Transposons

We used CENSOR software [25–27] to find and describe transposons in chloroplast genomes. That is a well-known fact that CENSOR works with the database of transposon patterns incorporated in it [25]. Thus, we searched for transposons in chloroplast genomes over the database provided by CENSOR and then used the transposons from CENSOR output to find them once more with convolution calculation. It should be stressed that we aim neither to search for new transposons in chloroplast genomes nor study their function in chloroplast genomes, but to compare two methodologies of symbol sequences comparison and a pattern search.

2.2.5. Intergenic DNA fragments from chloroplasts

We studied 101 chloroplast genomes of a hundred of species; *Monotropa hypopitys* was presented with two genomes of different organisms of the same species. It should be stressed that these two genomes are not identical at all, so we included both of them in the database. All genomes were retrieved from EMBL–bank. The database comprises the chloroplast genomes of 63 families; besides, we included cyanobacteria (two species) and protists (three species) into the data set. These latter are able to converse sunlight into a bioproduct, that is why they were included. This database contains four species of gymnosperm and 86 species of leaf plants.

The intergenic fragments of a genome were identified with the annotation and retrieved from the genome sequences with ad hoc software. We completely refer on the annotation, so that these former were defined as subsequences located in a genome between the end of j-th gene and the start of the $i+1$-th gene. On average, the number of intergenic fragments per genome was $\sim 10^2$ entities. Similar numbers were observed for cyanobacteria and protist genomes. The database contains 17 256 entries.

2.3. Elastic map

The elastic map technique is a powerful up-to-date tool for multidimensional data analysis and visualization. In brief, it implies approximation of multidimensional data with a manifold of low (typically of two) dimensions. A manifold is a core object of topology; here, we use a square from a Euclidean plane. An idea of approximation consists of proper deformation of a square to adjust the data best of all. Any transformation (expansion, squeezing, torsion, bending, etc.) is allowed except glue and discontinuity; in other words, the manifold's topology must remain.

An implementation of an elastic map starts from the determination of the first and the second principal components over the data set. A straight plane is developed then over these axes, and all the data should be projected on the plane. Then a minimal square comprising all the projections must be determined.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2022 doi:10.20944/preprints202201.0425.v1

At the next stage, each data point is connected to its projection with a mathematical spring; that latter has infinite tensility and remains the linear expansion rule. As soon as all the springs are erected, the originally rigid square must be changed for an elastic membrane. It must be uniform and homogeneous from the point of view of elasticity. Right now, the construction must be released to reach the minimum of the total deformation energy; the final deformation is unique and stable for the given data point configuration. The formal description of this part of the method could be found in [28,29].

Upon the final deformed of a manifold, the data point images on the jammed surface must be redefined: new images are determined as the orthogonal projections of the data points on the jammed surface. Practically, it means that the new image is the point on the deformed manifold closest to the original data point.

Finally, all the springs should be cut-off so that the jammed surface gets back into a stretch plane: this is the so-called *inner coordinate* representation. Of course, such inverse transformation of a jammed surface into a plane square modifies the position of images: this is the way to find out clusters if any. The images on a plane square defined in inner coordinates could be considered the points on the Euclidean plane; hence, many clustering methods could be implemented. We used the method based on local density determination.

To define local density, each point must be supplied with a bell-shaped function; we used a well-known Gaussian curve[1] $f_l(r)$:

$$f_l(r) = \exp\left\{-\frac{(r-r_l)^2}{\mu^2}\right\} \tag{5}$$

here index l enlists the points in the dataset, r_l is the location of l-th point on the plane, and μ is the contrast parameter. To define the local density, one must sum up all the functions (5) and depicture this sum function (6):

$$\mathcal{F}(r) = \sum_{l=1}^{M} f_l(r)\,. \tag{6}$$

Here M is the total number of points in the dataset. Function (6) level lines effectively identify clusters: these are the areas on the square with the $\mathcal{F}(r)$ value exceeding the given one.

2.4. Convolution in DNA sequence similarity search

To reveal a structure in DNA (or RNA) sequences, one has no other way but to compare nucleotide sequences. Currently, alignment is the leading method here. It has several crucial problems, including divergence, hard computations, arbitrariness in fine function choice, inability to go on with arbitrary long sequences. Insertions and deletions make the worst problem for alignment. Originally, alignment takes start from editing distance idea [30–33]. However, the popularity of alignment-free methods grow up, so we briefly present the novel one used to find transposons in chloroplast genomes.

The convolution-based method to compare nucleic sequences gathers together three fine simple ideas to address the problem. The first idea is to change the nucleotide sequence for a numeric one. A nucleotide sequence is to be changed for four (for each nucleotide separately) numeric sequences. Let $\mathfrak{T}$ be a nucleotide sequence; then $\mathcal{T}_A$ is derived from $\mathfrak{T}$ due to the substitution of 1s instead of A, and changing all other symbols for zero. Similarly, this procedure must be carried out for each of four nucleotides, so the four numeric sequences take place: $\mathcal{T}_\mathsf{A}$, $\mathcal{T}_\mathsf{C}$, $\mathcal{T}_\mathsf{G}$ and $\mathcal{T}_\mathsf{T}$. Obviously, the length of each binary sequence is equal to N (i. e. to the length of the original nucleotide sequence).

[1] It should be kept in mind that the Gaussian function here has nothing to do with normal distribution.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2022 doi:10.20944/preprints202201.0425.v1

The second idea is the core one: it stipulates the binary sequences to be the coefficients of a polynomial (of the degree $N-1$). So, a symbol-by-symbol comparison of two (symbol) sequences should be changed for a product (convolution, to be exact) of two polynomials. Finally, the third idea is to implement Fourier Transform for convolution calculation; moreover, Fast Fourier Transform should be used here to accelerate the calculations significantly.

The method works as follows. Suppose, $\mathfrak{T}_1$ and $\mathfrak{T}_2$ be the nucleotide sequences to be compared; let N_1 and N_2 be the lengths of each of them, correspondingly. Convert both of them into four binary sequences each; next, they both must be expanded to the length $M = 2^n \geq N_1 + N_2$, $M \mapsto \min$. The expansion is provided by adding as many zeros (from the right, for the sake of definiteness), as necessary to reach M.

Let now introduce the **convolution** $\mathbf{S} = \mathbf{A} \otimes \mathbf{B}$ of two number sequences $\mathbf{A} = \{a_i\}$, $i = 0, 1, 2, \ldots, N_1 - 1$ and $\mathbf{B} = \{b_i\}$, $i = 0, 1, 2, \ldots, N_2 - 1$. We may put $N_1 > N_2$ with no loss of generality. It is the sequence $\mathbf{S} = \{s_i\}$, $i = 0, 1, 2, 3, \ldots, N_1 + N_2 - 2$ with

$$s_i = \begin{cases} \sum\limits_{k=0}^{i} a_k b_{i-k}\,, & \text{if} \quad i < N_2\,, \\ \sum\limits_{k=0}^{N_2} a_k b_{i-k}\,, & \text{if} \quad N_2 \leq i < N_1\,, \\ \sum\limits_{k=0}^{N_1+N_2-1-i} a_{N-1-k} b_{i-N+1+k}\,, & \text{if} \quad N_1 \leq i\,. \end{cases} \tag{7}$$

The brute force way to calculate a convolution of two sequences is rather hard. To overcome this problem, we consider a convolution as a product of two polynomials (of the power $L-1$ and $N-1$, respectively). In other words, we consider two number sequences $\mathbf{A}$ and $\mathbf{B}$ as the sets of coefficients of corresponding polynomials. Thus, a convolution is converted to the product of two polynomials.

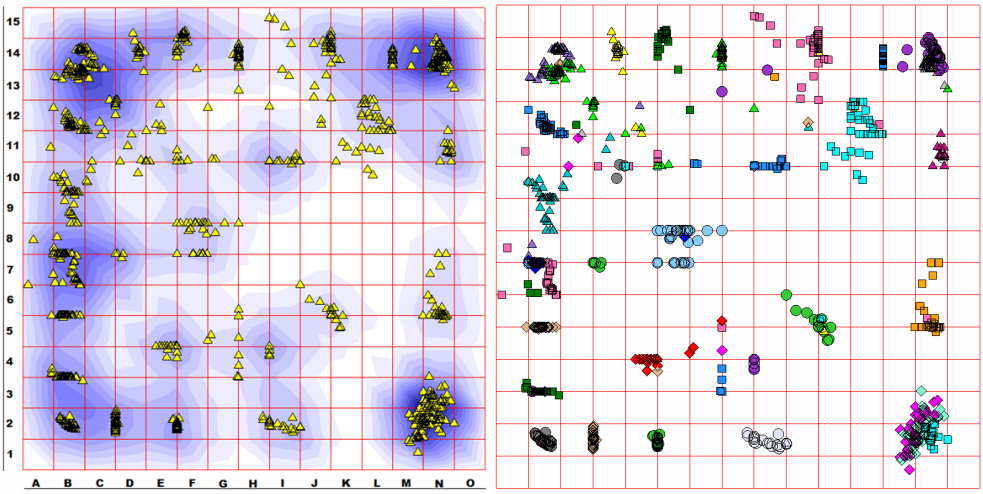
The next step comes from a well-known theorem saying that the Fourier transform of a convolution is the product of the Fourier transforms of the functions (sequences, in our case). Hence, the idea is to get the product of two polynomials with (fast) Fourier transform to both sequences, multiply the Fourier images, and apply the inverse Fourier transform to get the convolution of the original sequences. Fourier transform is, in turn, the convolution. Meanwhile, there is the specific algorithm of high-speed calculation of Fourier image of any number sequence called fast Fourier transform (FFT). Let $\mathbb{F}$ be FFT transforming a number sequence $\mathbf{A}$ into the sequence $\mathbf{A}' = \mathbb{F}(\mathbf{A})$ of the same length $N-1$. Let $\mathbb{F}^{-1}(\mathbf{A}') = \mathbf{A}$ be the inverse FFT. The operation $\mathbf{X} \boxtimes \mathbf{Y}$ for two number sequences $\mathbf{X} = \{x_i\}$, $i = 0, 1, 2, \ldots, N-1$ and $\mathbf{Y} = \{y_i\}$, $i = 0, 1, 2, \ldots, N-1$ yields:

$$\mathbf{X} \boxtimes \mathbf{Y} = \{x_i \cdot y_i\} \qquad i = 0, 1, 2, \ldots, N-1\,. \tag{8}$$

Consider two finite symbol sequences $\mathbf{P} = \{p_i\}$, $i = 0, 1, 2, \ldots, N_1 - 1$ and $\mathbf{Q} = \{q_i\}$, $i = 0, 1, 2, \ldots, N_2 - 1$ from alphabet $\aleph = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$. The algorithm comprises the following steps.

1. Inverse the sequence $\mathbf{Q}$ yielding $\widetilde{\mathbf{Q}} = \{q_{N_2-1-i}\}$ $i = 0, 1, 2, \ldots, N_2 - 1$.
2. Change $\mathbf{P}$ and $\widetilde{\mathbf{Q}}$ into $|\aleph| = 4$ binary sequences as described above.
3. Expand the sequences with zeros for further application of FFT to get the sequence of the length $N_1 + N_2 - 1$. To do it, all $2 \times |\aleph|$ binary sequences must be accomplished with zeros (to the right, for certainty) to that length. An effective implementation of FFT requires a sequence to have the length equal to power of 2, so we must to add zeros to get the length
$$\widetilde{N} = 2^{\lceil \log_2(N_1+N_2-1) \rceil}\,.$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2022 doi:10.20944/preprints202201.0425.v1



(a) Density distribution (b) Amino acids distribution

Figure 1. Clustering of 4,887 tRNA of 143 gymnosperm species. Local density distribution is shown in left (1(a)), with contrast radius $\mu = 0.15$. Total distribution of 22 amino acids is shown right (1(b)). Color scheme for amino acids is described in text. The elastic maps are shown in inner coordinates, see subsec. 2.3; we used 15 level scoring.

4. Apply FFT to each of the binary sequences:

$$\mathbf{P}'_{\mathsf{A}} = \mathbb{F}(\mathbf{P}_{\mathsf{A}}), \quad \mathbf{P}'_{\mathsf{C}} = \mathbb{F}(\mathbf{P}_{\mathsf{C}}), \quad \mathbf{P}'_{\mathsf{G}} = \mathbb{F}(\mathbf{P}_{\mathsf{G}}), \quad \mathbf{P}'_{\mathsf{T}} = \mathbb{F}(\mathbf{P}_{\mathsf{T}}),$$
$$\widetilde{\mathbf{Q}}'_{\mathsf{A}} = \mathbb{F}(\widetilde{\mathbf{Q}}_{\mathsf{A}}), \quad \widetilde{\mathbf{Q}}'_{\mathsf{C}} = \mathbb{F}(\widetilde{\mathbf{Q}}_{\mathsf{C}}), \quad \widetilde{\mathbf{Q}}'_{\mathsf{G}} = \mathbb{F}(\widetilde{\mathbf{Q}}_{\mathsf{G}}), \quad \widetilde{\mathbf{Q}}'_{\mathsf{T}} = \mathbb{F}(\widetilde{\mathbf{Q}}_{\mathsf{T}}).$$

5. Following (8), combine the relevant couples of $\mathbf{P}'_\nu$ and $\mathbf{Q}'_\nu$ with ν running A, C, G and T) and sum up them:

$$\mathbf{S}' = \mathbf{P}'_{\mathsf{A}} \boxtimes \mathbf{Q}'_{\mathsf{A}} + \mathbf{P}'_{\mathsf{C}} \boxtimes \mathbf{Q}'_{\mathsf{C}} + \mathbf{P}'_{\mathsf{G}} \boxtimes \mathbf{Q}'_{\mathsf{G}} + \mathbf{P}'_{\mathsf{T}} \boxtimes \mathbf{Q}'_{\mathsf{T}}.$$

6. Apply inverse FFT to $\mathbf{S}'$ thus getting the convolution $\mathbf{S} = \mathbb{F}^{-1}(\mathbf{S}')$.

3. Results

Here we present the effects revealed from the triplet composition of chloroplast genomes or their fragments. We start from the presentation of tRNA genes structure–function interplay, then change for symmetry presentation found in chloroplast genomes, then change for transposons analysis in chloroplast genomes.

3.1. T-genome and structure–function interplay

We analyzed 4,887 tRNA genes from 143 species of gymnosperms. The set of tRNA genes was identified for each species, and each gene was converted into W_3 triplet frequency dictionary (step $t = 1$). To cluster, a triplet must be excluded since the sum of all 64 frequencies makes 1. We excluded the triplet ACA from the analysis yielding the least standard deviation determined over the entire set of genes, $\sigma_{\mathsf{ACA}} = 0.00409$. To compare with, the largest figure is $\sigma_{\mathsf{GGT}} = 0.01055$. Table 2 shows the number of genes (all isodecoders are gathered in a group) for all 22 amino acids and initiating methionine. To find clusters, both linear (K-means) and non-linear (elastic maps) techniques have been implemented. No reliable classification has been observed with K-means; on the contrary, the elastic map technique yields an apparent cluster pattern in a soft 16×16 map.

Figure 1 shows the total distribution of all the genes involved in analysis over the 16×16 elastic map in inner coordinates. Local density distribution for $\mu = 0.15$ is shown in Subfig. 1(a), while the distribution of amino acids is shown in Subfig. 1(b). The clusters

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2022 doi:10.20944/preprints202201.0425.v1

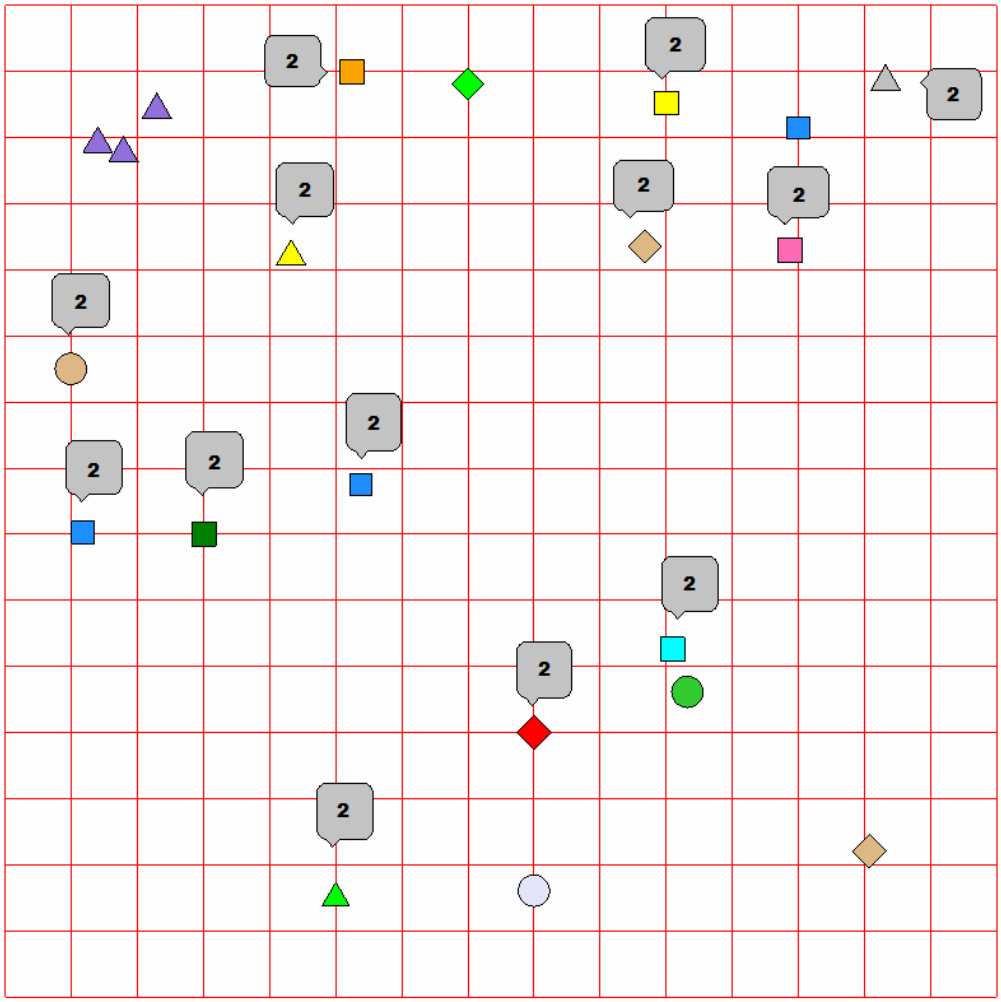


Figure 2. Individual distributions of the underrepresented tRNA genes; the insets show the number of genes in case they cover each other.

in the map are easily identified through the local density. We identified 15 clusters in Figure 1. The coloring scheme for amino acids presentation is shown in Table 1 (see subsec. 2.2 below), and the scheme is used systematically within the paper.

Fig. 1 contains 4 887 points labeled with four different marks, and colored in 21 color (see Table 1). Such mash presentation makes it difficult to see individual amino acid distributions. To make is easier, we provide a gallery of the individual amino acids distribution shown in Figs. 10 and 11. These figures show the distributions regardless of the synonymity; besides, the legends of subfigures show the number of synonym genes of each amino acid to make an easier analysis of the figures. The genes corresponding to individual amino acids in Figs. 10 and 11 (as well, as in Fig. 2) are shown in the location determined by the entire set of genes taken into consideration; we just erased the images of the genes other than those encoding the tRNAs transferring the specific amino acid. In other words, Fig. 1(b) is a combination of all the maps from Figs. 10 and 11.

The clusters in Fig. 1 are basically determined by the contrast parameter (correlation radius) μ; here we used $\mu = 0.15$ (see Eqs. (5, 6) in subsec. 2.3). The genes are distributed in triplet frequency space very inhomogeneously, forming clearly identified clusters; the question then arises whether the composition of the clusters is random or not, in terms of the specific genes comprised in a cluster. Fig. 1(b) answers this question: the clusters comprise the genes encoding tRNAs responsible for a specific amino acid transder. Figs. 10 and 11 also illustrate this fact clearly.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2022 doi:10.20944/preprints202201.0425.v1

Table 3. Number of copies of underrepresented genes of 12 amino acids; aa stands for amino acid, ω is a triplet, N is the abundance, RGB is the coloring scheme and $*$ is the label in Fig. 2.

aa	ω	N	RGB	$*$	aa	ω	N	RGB	$*$	aa	ω	N	RGB	$*$
A	AGC	2	255, 255, 0	△	K	AAA	3	147, 112, 219	△	T	CGU	2	0, 128, 0	□
	GGC	2	255, 255, 0	□	L	AUG	5	30, 144, 255	□	V	CAC	4	222, 184, 135	◇
C	ACA	2	255, 165, 0	□	P	CGG	2	0, 255, 255	□		AAC	2	222, 184, 135	○
D	AUC	2	255, 0, 0	◇	R	CCU	1	50, 205, 50	○	Y	AUA	2	192, 192, 192	△
E	TTC	1	230, 230, 250	○	S	AGA	2	0, 255, 0	△					
G	CCC	2	255, 105, 180	□		CGA	1	0, 255, 0	◇					

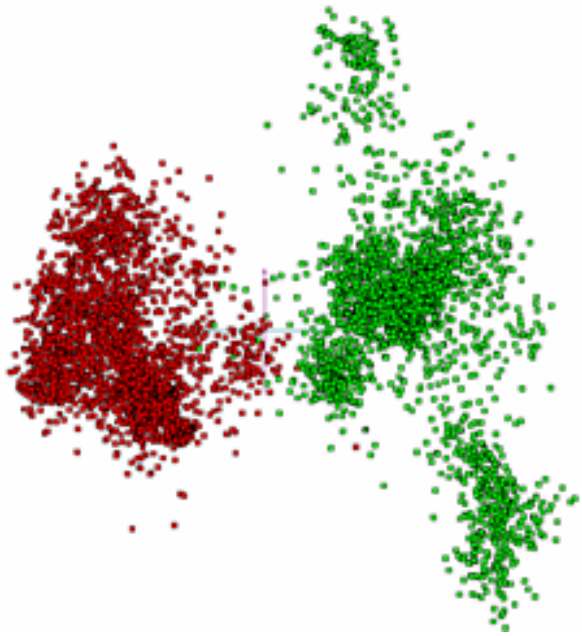


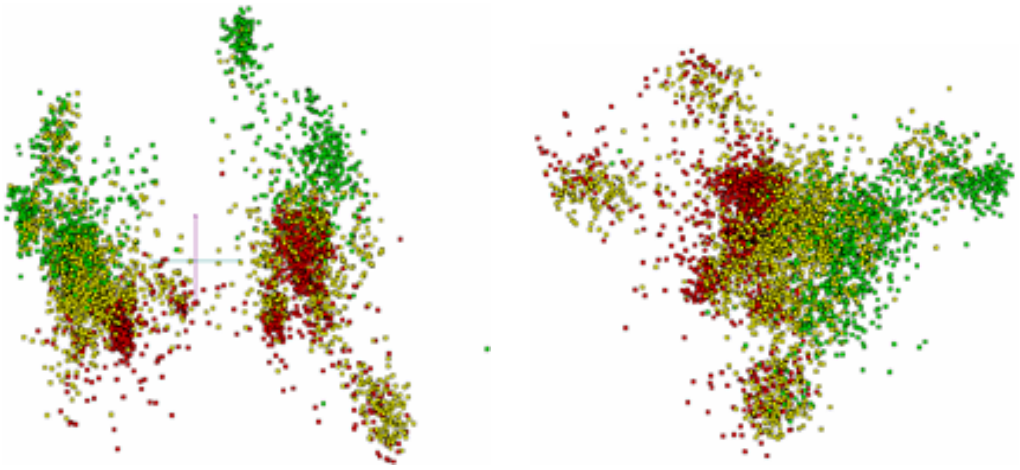
Figure 3. Photosystem genes shown in projection of the first and the second principal components plane. The genes from the forward strand are colored in red, the genes from the backward strand are colored in green.

The number of clusters in the subfigures shown in Figs. 10 and 11 exhibit a relatively low correlation to the number of synonyms. Indeed, the most significant majority of the gene families of the amino acids with two synonym codons has a single cluster; at least, there is no doubt that the cluster (if identified as two separated entities) heavily differs in an abundance of the genes. On the contrary, the numbers of genes with synonym codons are pretty close; for such genes, see Table 2.

The amazing fact is that the distribution of the isodecoders over the codons is very biased (cp. Table 2 and Figs. 10, 11). In other words, the number in Table 2 differs strongly if one splits them for the numbers of genes encoding different isodecoders. In fact, a single isodesoder is usually heavily underrepresented in a genome; only three amino acids fall beyond this pattern: alanine, serine, and valine. Table 3 shows the list of such gene types with underrepresented isodecoders; the distribution of them on an elastic map is shown in Fig. 2. However, it is unclear whether this bias naturally occurred in chloroplast genomes or resulted from the details of sequencing and/or annotation. Probably, one should expect the contribution from both factors.

The coloring scheme for amino acids in Table 3 is the same as in Table 1; however, the labels differ from Table 2 and mark the isodecoders. Fig. 2 shows the distribution of these underrepresented isodecoders over the elastic map (usually presented isodecoders are erased); as usual, the insets indicate the number of isodecoders with coincided images of the map. To avoid confusion, take a note that the marks of the same form and the same color indicate the different copies of the same isodecoder; e. g., lysine has three copies of the same (unique, in this case) isodecoder so they are shown as three triangles

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2022 doi:10.20944/preprints202201.0425.v1



(a) (PC_1, PC_2) (b) (PC_2, PC_3)

Figure 4. Distribution of genes with respect to CG-content, in (PC_1, PC_2) projection (Subfig. 4(a)) and in (PC_2, PC_3) projection (Subfig. 4(b)). Coloring scheme see in test.

colored in dark violet (RGB scheme [147, 112, 219]). The notation in this table is the same as in Table 1. Fig. 2 shows the distribution of these genes.

3.2. *Photosystem genes distribution*

We studied the distribution of W_3 dictionaries corresponding to the genes of photosystem located in chloroplast genomes. Fig. 3 shows this distribution in the projection on (PC_1, PC_2) plane determined by the first (PC_1) and the second (PC_2) principal components. The genes are gathered into apparent clusters; to begin with, we start from the distribution of the genes between the clusters in terms of their localization in the forward strand vs. backward one. Colorization shown in Fig. 3 unambiguously proves the separation of the corresponding genes between two distinct groups of clusters.

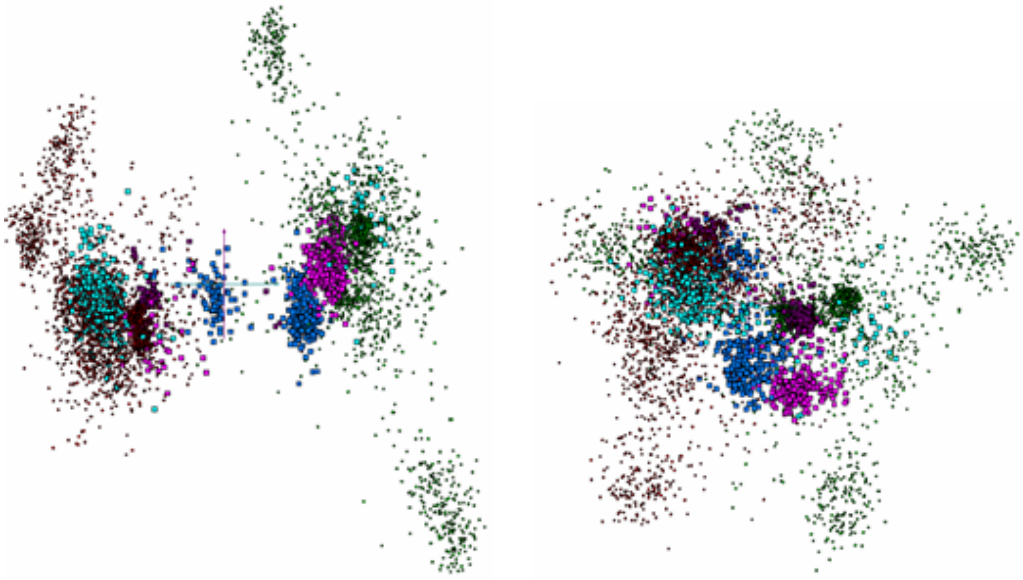
However, the question of whether the genes "prefer" to gather into clusters following the taxonomy or the function is not answered yet, with Fig. 3. To address the question, we checked the distribution of the genes belonging to the same type (e. g., *pca*B) in this figure. The genes of the same type are specifically colored. Doubtlessly, the genes comprise very dense clusters, and no taxonomy impact is seen in them; *psb*E genes are shown in rose carmine, *psb*K genes are shown in sky blue, *psa*B genes are shown in light red-violet, and *psa*C are shown in blue. Fig. 5 shows the distribution of four types of genes; we do not show the distribution for all the types since the number of types is high enough and no clear visualization could be achieved. However, we checked all the types and found similar patterns in genes distribution: all types of genes comprise very dense (sub)clusters in this pattern.

The clustering shown in Fig. 5 and similar patterns is very common for various genetic systems met in any taxonomy group [34,35]. Thus, a question towards the key factor identification determining such clustering arises; of course, standard (and advanced) statistical analysis techniques answer this question regarding the triplets contributing most of all to the separability of clusters. However, a generalized factor is more valuable and informative; CG-content is that latter. Indeed, this parameter is highly correlated to cluster pattern; Fig. 4 illustrates this fact. Here the genes colored in brown possess the highest CG-content level, and those colored in green exhibit the lowest one.

3.3. *ATP synthase and $NADH^+$ genes distribution*

We studied the distribution of ATP synthase, and $NADH^+$ genes in triplet frequency (Euclidean) space. Again, the gene level demonstrates an advantage of function over taxonomy: triplet frequency dictionaries exhibit robust clustering into very distinct and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2022 doi:10.20944/preprints202201.0425.v1



(a) (PC_1, PC_2) (b) (PC_2, PC_3)

Figure 5. Localization of the genes of different types within the core clusters, in (PC_1, PC_2) projection (Subfig. 5(a)) and in (PC_2, PC_3) projection (Subfig. 5(b)). Coloring scheme see in text.

apparent groups isolated from each other. The composition of these groups exhibits a very high correlation to a gene type, not to the taxonomy of the bearers of the genes.

Fig. 6 illustrates this fact. Subfig. 6(a) shows the distribution of ATP synthase genes shown in elastic map in so called inner coordinates; the genes belonging to different types are colored differently. The coloring scheme is as following: *atp*A is colored in magenta, *atp*B is colored in green, *atp*E is colored in dark orange, *atp*F is colored in dark violet, *atp*H is colored in light pink and *atp*I is colored in brown. All genes are labeled with circles in this Fig., regardless their type or taxonomy.

Subfig. 6(b) shows the distribution of the most abundant taxa in the same map. The genes are labeled with triangles, with the following coloring scheme: *Fabaceae* family is colored in red, *Malvaceae* family is colored in lime and *Brassicaceae* family is colored in yellow. Immediate comparison of these two subfigures unambiguously proves the total prevalence of function (i. e. type of a gene) over the taxonomy of its bearer.

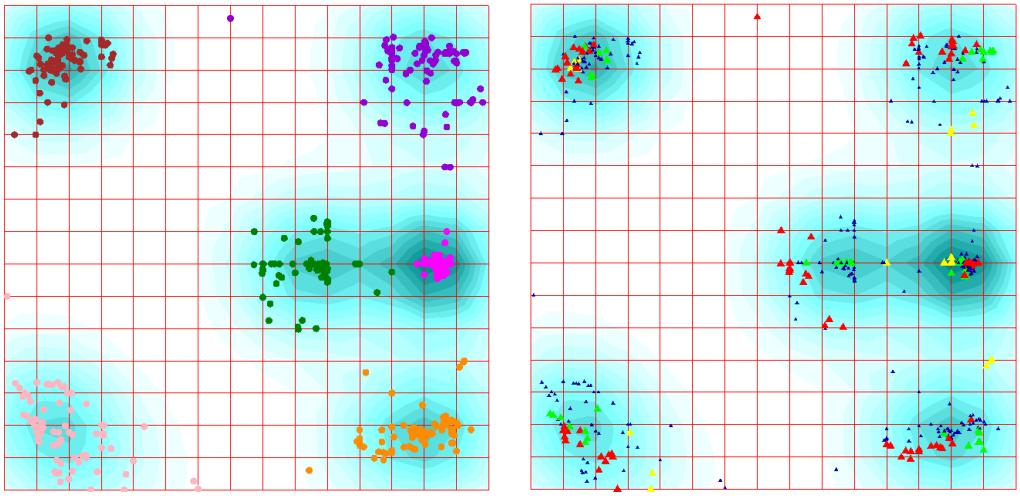
Similarly, Fig. 7 shows the distribution of $NADH^+$ genes located in chloroplasts over the elastic map in inner coordinates. There are eleven types of these genes with the following coloring scheme: *ndh*A is colored in light pink, *ndh*B is colored in brown, *ndh*C is colored in gray, *ndh*D is colored in black, *ndh*E is colored in coral, *ndh*F is colored in lime, *ndh*G is colored in pink, *ndh*H is colored in yellow, *ndh*I is colored in yellow green, *ndh*J is colored in plum and *ndh*K is colored in red.

Subfig. 7(a) shows the distribution of eleven types of $NADH^+$ genes over the elastic map in inner coordinates. All types of genes are colored as described above. On the contrary, Subfig. 7(b) shows the distribution of the most abundant taxa over the same map; evidently, the taxons are spread almost equally among the clusters, unlike the functional types of genes.

3.4. Stratification of intergenic fragments of chloroplast genomes

We run the same procedure for clustering, as for tRNA, ATP synthase, and $NADH^+$ genes. Surely, the difference in the inner structure of triplet composition mentioned in subsec. 1.5 significantly affects the results. Since clustering for intergenic fragments of genomes is quite fuzzy, we used the maps of various elasticity to analyze the distribution patterns. Fig. 8 shows the distribution of intergenic fragments of all studied chloroplast genomes; Subfig. 8(a) shows the distribution developed on elastic map of 16×16 size

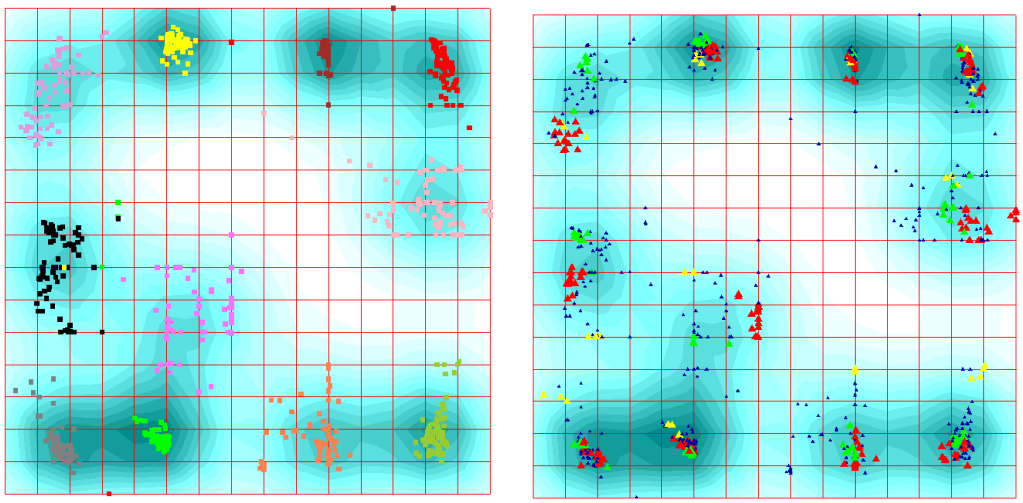
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2022 doi:10.20944/preprints202201.0425.v1



(a) (b)

Figure 6. Distribution of ATP synthase genes located in chloroplasts. Subfig. 6(a) shows the distribution of gene types, and Subfig. 6(b) shows the distribution of taxons.

(so-called *soft map*), and Subfig. 8(b) shows similar distribution developed on elastic map of 25 × 25 size (so-called *detailed map*); local density is shown in cyan.



(a) (b)

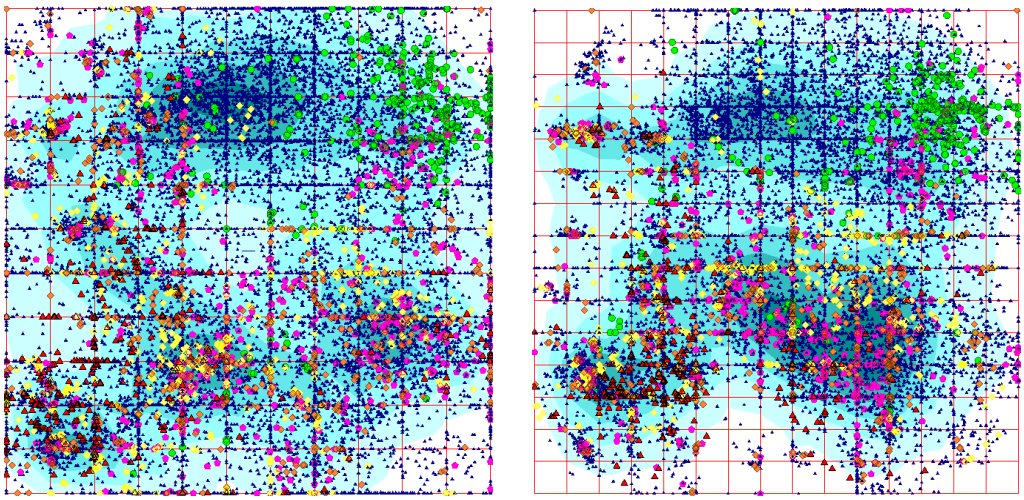
Figure 7. Distribution of $NADH^{+}$ genes located in chloroplasts. Subfig. 7(a) shows the distribution of gene types, and Subfig. 7(b) shows the distribution of the most abundant taxons.

Five most abundant taxa are colored in this Figure: green circles label *Marchantiophyta* order, red triangles label *Pinaceae* family, pink circles label *Rosaceae* family, orange triangles label *Solanaceae* family, and finally yellow rhombae label *Oryza* family. All other species are shown in dark blue small triangles.

First of all, a similar clustering pattern is observed in both maps, regardless of its elasticity. Remarkably, *Marchantiophyta* order genes produce clearly separated clusters (cluster is located in the upper right corner of the maps). Besides, distinct clusters gather cpDNA fragments that is clearly seen in the soft map (6 × 16).

Also, a distribution of intergenic fragments retrieved from the same genome is of interest. We studied such distribution, as well. Subfig. 9(a) shows this distribution. Besides, the database contains four genomes of non-photosynthetic species; Subfig. 9(b) shows the distribution of these entries on the elastic map. It should be stressed that all

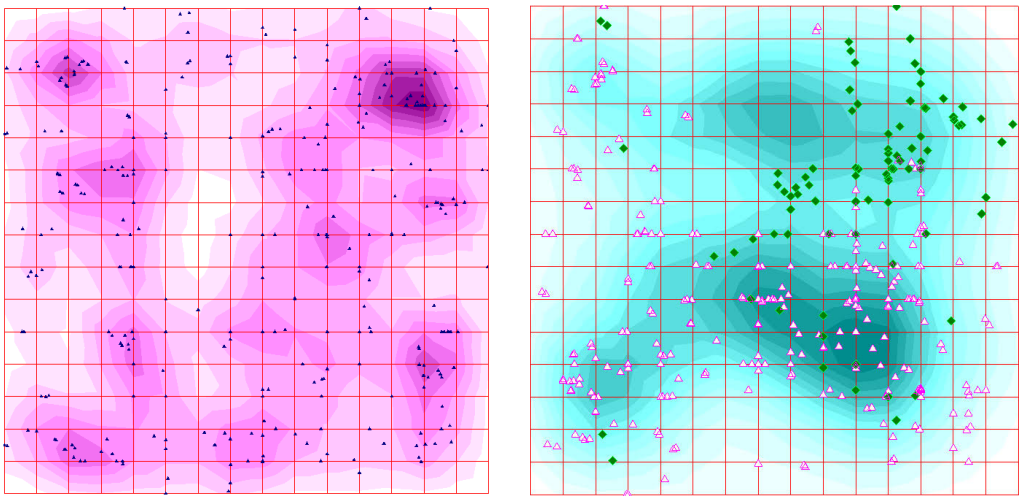
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2022 doi:10.20944/preprints202201.0425.v1



(a) (b)

Figure 8. Distribution of intergenic fragments in triplet frequency space, see coloring scheme in text. Subfig. 8(a) shows rigid map (12 × 12), Subfig. 8(b) shows the soft map (16 × 16).

images of other genomes are erased in this figure. One can see that the general local density of the points distribution in this figure is the same as in Fig. 8.



(a) (b)

Figure 9. Distribution of intergenic fragments of *Rosaceae* family (Subfig. 9(a)) and distribution of intergenic fragments of four genomes of non-photosynthetic species (Subfig. 9(b)

3.5. Convolution approach to a pattern search

Let us now illustrate the power of the convolution-based method of a pattern search with transposons to be found in chloroplast genomes. We compared two methodologies of a pattern search: the former is the alignment-based search provided by CENSOR software, and the latter is convolution methodology. To begin with, CENSOR has its own database of transposons; however, it is out of public access. Meanwhile, CENSOR output contains the transposons sequences found in genomes, so we used these found sequences for convolution search.

Table 4 shows the results of the comparison of the execution of software for these two methods. The table shows the results for gymnosperms only. For the sake of brevity, this table shows the compressed data for families level only. N is the number of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2022 doi:10.20944/preprints202201.0425.v1

Table 4. Comparison of CENSOR (rigid filtering) vs. convolution search results obtained on gymnosperm chloroplast genomes; see explanation in text.

Transposon	CENSOR			Convolution		
	N	$\langle S \rangle$	$\sigma_{\langle S \rangle}$	N	$\langle L \rangle$	$\sigma_{\langle L \rangle}$
Pinaceae, 56 species						
Copia-18_BD-I	56	248,2	2,3	57	295,0	1,1
MuDR-64_OS				29	281,3	8,0
Araucariaceae, 4 species						
Copia-18_BD-I	4	259,3	2,9	4	298,8	1,0
Taxaceae, 4 species						
Copia-18_BD-I	4	236,0	3,5	4	291,0	1,2
MuDR-64_OS				1	270,0	0,0
Cycadaceae, 5 species						
Copia-18_BD-I	10	263,0	0,0	10	300,0	0,0
Zamiaceae, 7 species						
Copia-18_BD-I	7	262,6	1,1	7	299,9	0,3
MuDR-64_OS				3	283,3	6,9
Podocarpaceae, 3 species						
Copia-18_BD-I	3	234,0	1,7	3	262,7	0,6
Ephedraceae, 4 species						
Copia-18_BD-I	8	233,0	0,0	8	290,0	0,0
MuDR-64_OS				1	249,0	0,0
Ginkgoaceae, 1 species						
Copia-18_BD-I	2	260,0	0,0	2	299,0	0,0

transposon entries in a family for Copia18_BD-I and MuDR-64_OS, separately. $\langle S \rangle$ is the score of alignment averaged over a family; $\langle L \rangle$ is the transposon length as determined with convolution technique, correspondingly. Finally, $\sigma_{\langle S \rangle}$ and $\sigma_{\langle L \rangle}$ are the standard deviation of the score and length, respectively. Similarly, Tables 5 and 6 show the same figures observed on loaf plant chloroplast genomes.

To begin with, CENSOR allows two options in parameters choice: rigid filtration and soft filtration. Maybe, reconfiguration may result in another parameter set; however, this opportunity is not obvious for a user. CENSOR shows an occurrence of MuDR-64_OS transposon among gymnosperms obviously differs from that in loaf plants; however, the situation differs strongly if one changes the rigid filtration for a soft one. The point is that soft filtration brings much stuff in output, i. e. the sequences which are not transposons. Moreover, the average length of the sequences claimed to be transposons grows heavily, for soft filtration reaches the figure of $1,500$ nucleotides. It may result from the specific parameters set choice, which yields an excessive growth of various types of mismatches including insertions or deletions.

On the contrary, convolution yields the number of exactly matching nucleotides regardless of their location along with a pattern (transposon in our case).

4. Discussion

We studied various specific genetic systems (photosystem genes, tRNA genes, etc.) with the similar investigation tool: that is a triplet frequency dictionary, and the patterns they yield in triplet frequencies Euclidean space. General questions for all the issues here are following:

Q 1: do the genetic entities under consideration tend to gather into clusters, and

Q 2: what is the leading factor determining the composition of those clusters, if any?

Here a number of examples of cluster structures observed over the entities is provided. Generally speaking, the following "theorem" holds true:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2022 doi:10.20944/preprints202201.0425.v1

Table 5. CENSOR vs. convolution competition, the denotations are the same as in Table 4.

transposon	N	$\langle S \rangle$	$\sigma\_\langle S \rangle$	transposon	N	$\langle L \rangle$	$\sigma\_\langle L \rangle$
Caprifoliaceae, 2 species							
MuDR-64_OS	4	337.0	10.1	$MuDR-64\_OS$	4	538.5	6.4
Copia-18_BD-I	8	272.0	0.0	$Copia-18\_BD-I$	8	303.0	0.0
Oleaceae, 1 species							
MuDR-64_OS	2	345.0	0.0	$MuDR-64\_OS$	2	565.0	0.0
Copia-18_BD-I	2	269.0	0.0	$Copia-18\_BD-I$	2	302.0	0.0
Malvaceae, 1 species							
Copia-18_BD-I	2	278.0	0.0	$Copia-18\_BD-I$	2	305.0	0.0
				$MuDR-64\_OS$			
Fabaceae, 2 species							
MuDR-64_OS	2	300.0	5.7	$MuDR-64\_OS$	2	499.0	5.7
Copia-18_BD-I	4	275.0	0.0	$Copia-18\_BD-I$	4	304.0	0.0
Sapindaceae, 23 species							
MuDR-64_OS	46	329.9	1.4	$MuDR-64\_OS$	46	561.1	1.1
Copia-18_BD-I	46	281.1	0.6	$Copia-18\_BD-I$	46	306.0	0.0
Berberidaceae, 1 species							
MuDR-64_OS	2	394.0	0.0	$MuDR-64\_OS$	2	555.0	0.0
Copia-18_BD-I	2	275.0	0.0	$Copia-18\_BD-I$	2	304.0	0.0
Asteraceae, 1 species							
MuDR-64_OS	2	344.0	0.0	$MuDR-64\_OS$	2	558.0	0.0
Copia-18_BD-I	2	272.0	0.0	$Copia-18\_BD-I$	2	303.0	0.0
Amaranthaceae, 3 species							
				$MuDR-64\_OS$	12	292.3	0.5
Copia-18_BD-I	6	249.0	3.5	$Copia-18\_BD-I$	12	280.2	16.6
Poaceae, 1 species							
MuDR-64_OS	1	577.0	0.0	$MuDR-64\_OS$	2	599.0	0.0
Copia-18_BD-I	1	310.0	0.0	$Copia-18\_BD-I$	2	316.0	0.0
Copia-33_BD-I	2	595.0	0.0				
Helitron-N17B_OS	2	229.0	0.0				
Chrysobalanaceae, 1 species							
MuDR-64_OS	2	289.0	0.0	$MuDR-64\_OS$	2	521.0	0.0
Copia-18_BD-I	2	272.0	0.0	$Copia-18\_BD-I$	2	303.0	0.0
Nyctaginaceae, 1 species							
				$MuDR-64\_OS$	2	287.0	0.0
Copia-18_BD-I	2	263.0	0.0	$Copia-18\_BD-I$	2	300.0	0.0
Solanaceae, 2 species							
MuDR-64_OS	4	340.0	0.0	$MuDR-64\_OS$	4	535.0	0.0
Copia-18_BD-I	4	270.5	2.1	$Copia-18\_BD-I$	4	302.5	0.7
Ranunculaceae, 19 species							
MuDR-64_OS	30	335.5	5.2	$MuDR-64\_OS$	30	532.7	2.8
Copia-18_BD-I	30	275.0	0.0	$Copia-18\_BD-I$	30	304.0	0.0
Acoraceae, 2 species							
MuDR-64_OS	4	386.0	0.0	$MuDR-64\_OS$	4	577.0	0.0
Copia-18_BD-I	4	275.0	0.0	$Copia-18\_BD-I$	4	304.0	0.0
Actinidiaceae, 12 species							
Copia-18_BD-I	24	313.0	0.0	$Copia-18\_BD-I$	24	304.0	0.0
				$MuDR-64\_OS$	11	586.4	13.1

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2022 doi:10.20944/preprints202201.0425.v1

Table 6. CENSOR vs. convolution competition, the denotations are the same as in Table 4.

transposon	N	$\langle S \rangle$	$\sigma\_\langle S \rangle$	transposon	N	$\langle L \rangle$	$\sigma\_\langle L \rangle$
Adoxaceae, 1 species							
MuDR-64_OS	2	357.0	0.0	$MuDR-64\_OS$	2	558.0	0.0
Copia-18_BD-I	2	263.0	0.0	$Copia-18\_BD-I$	2	300.0	0.0
Amaryllidaceae, 39 species							
MuDR-64_OS	74	294.6	11.1	$MuDR-64\_OS$	75	520.2	15.7
Copia-18_BD-I	80	283.6	1.1	$Copia-18\_BD-I$	80	306.9	0.4
Akaniaceae, 1 species							
Copia-18_BD-I	2	275.0	0.0	$Copia-18\_BD-I$	2	304.0	0.0
Alstroemeriaceae, 1 species							
MuDR-64_OS	2	347.0	0.0	$MuDR-64\_OS$	2	553.0	0.0
Copia-18_BD-I	2	250.0	0.0	$Copia-18\_BD-I$	2	277.0	0.0
Altingiaceae, 1 species							
MuDR-64_OS	6	350.0	0.0	$MuDR-64\_OS$	6	556.7	0.6
Copia-18_BD-I	6	284.0	0.0	$Copia-18\_BD-I$	6	307.0	0.0

All the entities considered in this study do provide a cluster pattern with distinct and clearly seen clusters. Function is the leading factor determining the composition of the clusters.

First of all, we should stress that the change of the consideration for a genome level form a gene one results in complete inversion of the leading factor: taxonomy of genome bearers almost perfectly matches the clusters composition [36].

Let now discuss the phenomenon of the function prevalence in genes distribution in the space of triplet frequencies and the symmetry in chloroplast genomes in more detail; see subsecs. 4.1, 4.2, 4.3 and 4.4). Also, we focus on a convolution-based approach to sequence comparison and pattern search. Let now discuss each item in more detail.

4.1. Clustering of photosystem genes from chloroplasts

Chloroplast genetic system is significantly integrated with the nuclear genome. It makes a study of the pattern of genes distribution representing a chloroplast part of the relevant genetic system quite hot. The results shown above unambiguously prove the substantial prevalence of function over taxonomy when a clustering structure is analyzed. This prevalence is understood as a leading role of the function encoded in the genes under consideration in the determination of the composition of a cluster.

The structure of a pattern of distribution of chloroplast genes of the photosynthetic systems observed in triplet frequency space differs to some extent from similar ones previously revealed for complete genomes of chloroplasts, mitochondria, and bacteria [36,37]. This difference may result both from taxonomy impact and from fine unique peculiarities of function. To address this question, one should figure out whether a generalized factor exists ruling the difference. CG-content looks to be a good candidate for this parameter; moreover, its efficiency in the determination of a triplet distribution structure peculiarities was approved for bacteria genomes [38–40].

Here we present some preliminary results of the investigation of a relation between population triplet structure composition and function of photosystem genes. One sees two ways to address the problem:

I. To extend the database of the organisms under investigation incorporating the plants from very far and divergent systematic groups, and

II. To compare the chloroplast genes of a photosystem to the non-chloroplast genes of the same or pretty close function.

At first glance, the first point is just a matter of time. However, that is not so. The growth of sequenced genetic data does not follow the natural diversity of various plant taxa found in nature. Thus, one may pursue two competing strategies here: the former

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2022 doi:10.20944/preprints202201.0425.v1

consists in the immediate analysis of all genomes as soon as they appear in genetic banks; the latter implies a selection of the species in the way the selected entities represent the natural diversity as good, as possible. Both approaches still await a researcher.

4.2. Clustering tRNA genes

Transfer RNA genes are another genetic object meeting the constraints for a study of the interplay between structure, function, and taxonomy. Again, triplet composition counted within a gene nucleotide sequence is the structure here. A function is well-defined: these genes encode the RNAs transferring the specific amino acids to a ribosome where the protein assembling runs. All other functions of these genes products is neglected, at the moment. So, the function difference here is stipulated in manifestation of the specificity in transfer of a specific amino acid residue; moreover, this function of these genes seems to be highly homogeneous over a great variety of taxa. Of course, other functions of these genes may affect the pattern we are discussing; however, the results shown above unambiguously approve the leading role of a type of amino acid residue transfer, in cluster pattern arrangement.

Clustering of a family of tRNAs exhibits a high-quality distinguishability of the genes (see Figs. 1 and 2). These figures show the distribution of tRNA genes for gymnosperms; practically the same distribution is observed for higher plants. Let now focus on Figs. 10 and 11. The key question here is the behaviour of isodecoders of amino acid. As one can see from Fig. 10, alanine, cisteine, glutamic acid, phenilalanine, lysine, asparagine, glutamine, and tyrosine exhibit quite unimodal distribution in inner coordinates: the greatest majority of genes comprise the main cluster, and the number of escapees is small.

On the contrary, glycine, isoleucine, lysine, proline, arginine, threonine, and valine have distributions with two isolated clusters. Other genes have distributions of quasi-diffusive type where the genes are spread over the map around the main cluster. Such divergence in the distribution patterns may reflect some fine differences in structure coming, in turn, from the difference in function. However, this issue requires further studies.

In general, tRNA genes tend to gather into separate clusters following the amino acid residue transfer specificity. We checked whether the species (families, to be exact) are spread among the clusters in some order: the answer is negative. There is no order or a preference in taxa distribution over the clusters. In other words, each sufficiently abundant cluster comprises the list of species pretty close to that one observed in another one. However, the rare escapees (i. e. the genes located solely far from the "mother" cluster) seem to be not randomly appeared entities. This might be a manifestation of the function of tRNA other than amino acid residue transfer; however, this assumption is waiting further studies. Anyway, a coarse pattern of the distribution of tRNA genes in the space of triplet frequencies unambiguously proves the leading role of function in clustering and cluster content.

4.3. ATP synthase and $NADH^+$ genes

Two more groups of genes retrieved from chloroplast genomes come from energy consumption complex: these are ATP synthase and $NADH^+$ genes. They also are known for their conservative evolution. Hence, these groups of genes provide a good genetic matter for the study of the interplay between structure, function, and taxonomy or bearers.

The distribution of ATP synthase genes ($NADH^+$ genes, respectively) is shown in Fig. 6 (7, respectively). Both groups show highly perfect clustering that follows the function of genes, not the taxonomy of the bearers. We do not show the distribution of other families since they are presented with a small number of species; here, bias may occur resulting from this number minority. However, the distributions showed in Figs. 6 and 7 demonstrate the strong prevalence of the function over taxonomy.

Remarkably, the "quality" of distribution here is much better than that latter observed for tRNA genes. Probably, this fact from function choice of a genes family: similarly high "quality" distribution was observed over ATP synthase genes family located in fungal mitochondria [41,42].

The obvious next step in the investigation of the distribution of these genes may consist in a study of the combined distribution of ATP synthase genes, and $NADH^+$ genes; regardless of the specificity of their localization in genomes. These two families of genes seem to be very good for such kind study, Of course, there might be (and probably are) some minor differences between, say, $NADH^+$ chloroplast genes of species A and those of species A. However, their functional proximity is strong enough to neglect it at the first step. Same is true for ATP synthase genes families. These two families (ATP synthase and $NADH^+$ genes from chloroplast genomes) are the only two sets to be found simultaneously in chloroplasts and mitochondria. So, a mutual distribution of the genes belonging to these two families brings further progress in understanding the interplay between function and taxonomy. Speaking in advance, we tried this mutual distribution and found the "double-level" pattern: the genes of specific types both of ATP synthase and $NADH^+$ families form very clear and discrete clusters. However, the clusters are separated over the elastic map into two discrete non-overlapping subsets. In other words, both taxonomy and function contribute almost equally to the distribution of genes; however, this point requires further investigations, and the detailed discussion of it falls beyond the scope of this paper.

4.4. Intergenic chloroplast genomes fragments

intergenic fragments of any genome are well-known to be different from the coding ones. First of all, they do not code any protein. Another essential difference consists in the statistical properties of alternation of nucleotides in a sequence. Consider frequency dictionary $W_{(3,3)}$ of triplets counted along a sequence with the step $t = 3$. In fact, there are three dictionaries differing in the starting position of a reading frame: indeed, since the reading frames do not overlap, there may be three starting positions for that former. Hence, one sees three frequency dictionaries $W^0_{(3,3)}$, $W^1_{(3,3)}$ and $W^2_{(3,3)}$ differing in the location of the starting nucleotide. So, the key issue is that the set of these three dictionaries counted over coding regions differs significantly from the similar set counted over a intergenic region.

This difference stands behind the methodology of identification of coding regions in newly sequenced genomic entities. Papers [34,38–40] report the effects resulting from this difference. On the contrary, intergenic regions do not yield a three-beam cluster structure. Nonetheless, they are an important part of any genome that may not be eliminated. Skipping the functional role of intergenic fragments, consider the interplay between the structure (that is triplet frequency dictionary $W_{(3,1)}$ counted with the step $t = 1$) and taxonomy of the genome bearer. There is probably no surprise to meet the taxonomy's impact on clustering structure if the triplet dictionaries represent several different organisms [43]. Briefly speaking, one can see the non-equilibrium distribution of the dictionaries in triplet frequency space, and this former has clusters. The composition of the clusters is also non-random: they comprise the dictionaries preferably belonging to the same taxon. However, this preference could be observed for sufficiently distant (whatever this distance might mean) taxa.

Author Contributions: Conceptualization, M.Sad. and Yu.P.; methodology, M.Sad., Yu.P. and M.Sen. statistical analysis and data retrieval, O.M., A.F., T.Sh. Ya.N. and M.Sen., software, computations and visualization, A.M. writing, M.Sad. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: Anna Molyavko has received the support from the Ministry of Science and Higher Education of the Russian Federation through Mathematical Center at Krasnoyarsk

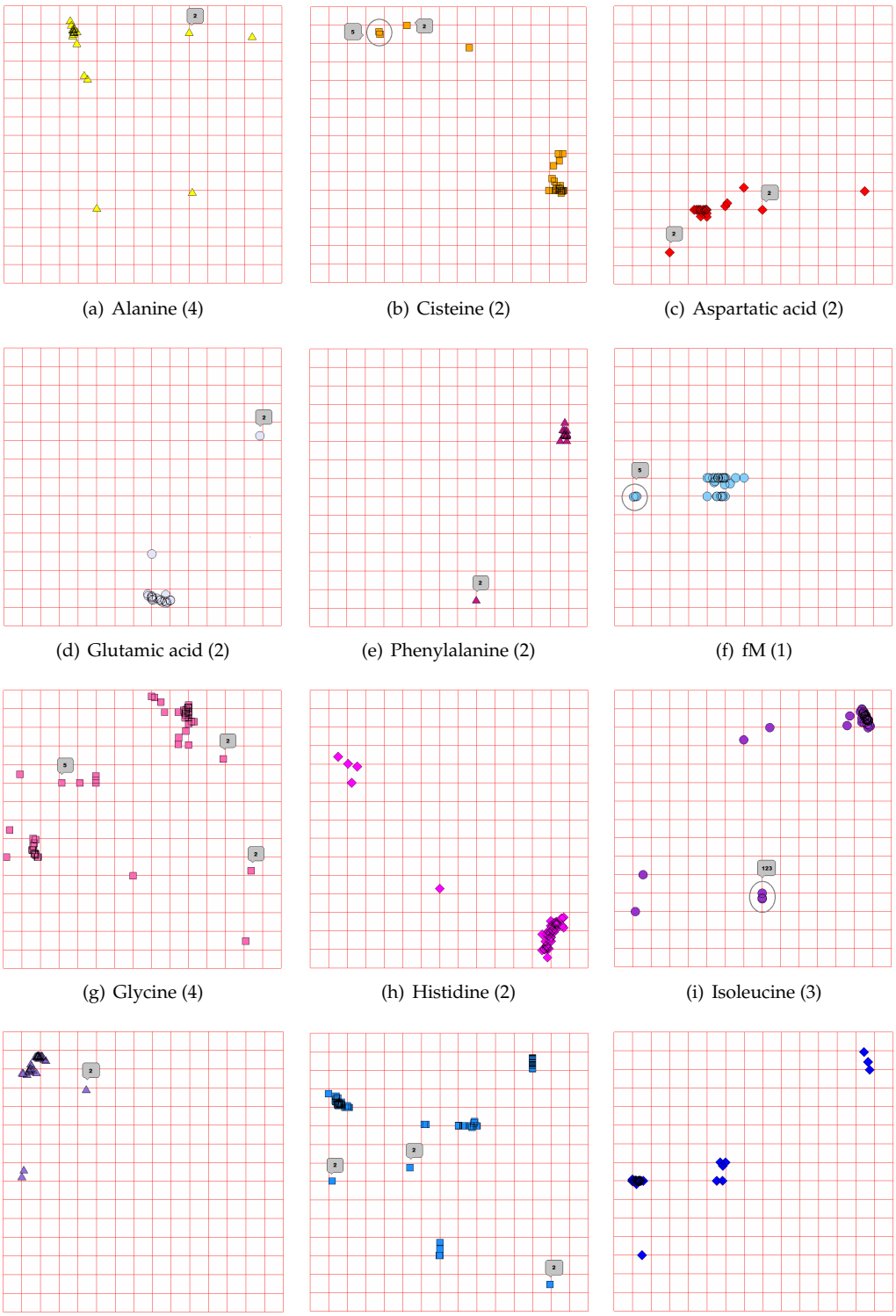
 doi:10.20944/preprints202201.0425.v1

(Agreement No. 075-02-2020-1631). This work was partly supported by Laboratory of forest genomics of Siberian federal university. Also, we are thankful toVladimir Shaidurov and Evgenia Karepova from ICM SB RAS and Igor Borovikov from Electronic Art, inc., California, for valuable discussion of convolution methodology.

Conflicts of Interest: The authors declare no conflict of interest.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2022 doi:10.20944/preprints202201.0425.v1

5. tRNA distribution pattern. Illustration



(a) Alanine (4) (b) Cisteine (2) (c) Aspartatic acid (2)

(d) Glutamic acid (2) (e) Phenylalanine (2) (f) fM (1)

(g) Glycine (4) (h) Histidine (2) (i) Isoleucine (3)

(j) Lysine (2) (k) Leucine (6) (l) Methionine (1)

Figure 10. Individual distributions of the specific tRNA genes; the insets show the number of genes in case they cover each other.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 January 2022 doi:10.20944/preprints202201.0425.v1

(a) Asparagine (2) (b) Proline (4) (c) Glutamine (2)

(d) Arginine (4) (e) Serine (4) (f) Threonine (4)

(g) Valine (4) (h) Tryptophan (1) (i) Tyrosine (2)

Figure 11. Individual distributions of the specific tRNA genes; the insets show the number of genes in case they cover each other.

References

1. Lanciano, S.; Cristofari, G. Measuring and interpreting transposable element expression. *Nature Reviews Genetics* **2020**, pp. 1–16.
2. Navarro, C. The mobile world of transposable elements. *Trends in Genetics* **2017**, *33*, 771–772.
3. Chan, K.X.; Phua, S.Y.; Crisp, P.; McQuinn, R.; Pogson, B.J. Learning the languages of the chloroplast: retrograde signaling and beyond. *Annual review of plant biology* **2016**, *67*, 25–53.
4. Zhang, X.; Bauman, N.; Brown, R.; Richardson, T.H.; Akella, S.; Hann, E.; Morey, R.; Smith, D.R. The mitochondrial and chloroplast genomes of the green alga Haematococcus are made up of nearly identical repetitive sequences. *Current Biology* **2019**, *29*, R736–R737.
5. Seidl, M.F.; Thomma, B.P. Transposable elements direct the coevolution between plants and microbes. *Trends in Genetics* **2017**, *33*, 842–851.
6. Wicker, T.; Gundlach, H.; Spannagl, M.; Uauy, C.; Borrill, P.; Ramírez-González, R.H.; De Oliveira, R.; Mayer, K.F.; Paux, E.; Choulet, F. Impact of transposable elements on genome structure and evolution in bread wheat. *Genome biology* **2018**, *19*, 1–18.
7. Kohl, S.; Bock, R. Transposition of a bacterial insertion sequence in chloroplasts. *The Plant Journal* **2009**, *58*, 423–436.
8. Huang, C.Y.; Ayliffe, M.A.; Timmis, J.N. Direct measurement of the transfer rate of chloroplast DNA into the nucleus. *Nature* **2003**, *422*, 72–76.
9. Grechishnikova, D.; Poptsova, M. Conserved 3′ UTR stem-loop structure in L1 and Alu transposons in human genome: possible role in retrotransposition. *BMC genomics* **2016**, *17*, 1–17.
10. Kojima, K.K. Human transposable elements in Repbase: genomic footprints from fish to humans. *Mobile DNA* **2018**, *9*, 2.
11. Fedoroff, N. Transposons and genome evolution in plants. *Proceedings of the National Academy of Sciences* **2000**, *97*, 7002–7007.
12. Haberle, R.C.; Fourcade, M.L.; Boore, J.L.; Jansen, R.K. Complete chloroplast genome of *Trachelium caeruleum*: extensiverearrangements are associated with repeats and tRNAs. *Journal of Molecular Evolution* **2006**, *66*.
13. Cullis, C.A.; Vorster, B.J.; Van Der Vyver, C.; Kunert, K.J. Transfer of genetic material between the chloroplast and nucleus: how is it related to stress in plants? *Annals of Botany* **2009**, *103*, 625–633.
14. Xiao-Ming, Z.; Junrui, W.; Li, F.; Sha, L.; Hongbo, P.; Lan, Q.; Jing, L.; Yan, S.; Weihua, Q.; Lifang, Z.; others. Inferring the evolutionary mechanism of the chloroplast genome size by comparing whole-chloroplast genome sequences in seed plants. *Scientific reports* **2017**, *7*, 1–10.
15. Bennetzen, J.L. Transposable element contributions to plant gene and genome evolution. *Plant molecular biology* **2000**, *42*, 251–269.
16. Costa, F.F. *Non-coding RNAs, epigenetics and complexity in human cells*; Vol. 7, Chapter, 2012.
17. Ludwig, M.Z. Functional evolution of noncoding DNA. *Current opinion in genetics & development* **2002**, *12*, 634–639.
18. Cobb, J.; Büsst, C.; Petrou, S.; Harrap, S.; Ellis, J. Searching for functional genetic variants in non-coding DNA. *Clinical and experimental pharmacology & physiology* **2008**, *35*, 372–375.
19. Balakirev, E.S.; Ayala, F.J. Pseudogenes: are they "junk" or functional DNA? *Annual review of genetics* **2003**, *37*, 123–151.
20. Cheong, E.J.; Myong-Suk, C.; Kim, S.C.; Kim, C.S. Chloroplast noncoding DNA sequences reveal genetic distinction and diversity between wild and cultivated *Prunus yedoensis*. *Journal of the American Society for Horticultural Science* **2017**, *142*, 434–443.
21. Tang, P.; Xu, Q.; Shen, R.; Yao, X. Phylogenetic relationship in *Actinidia* (*Actinidiaceae*) based on four noncoding chloroplast DNA sequences. *Plant Systematics and Evolution* **2019**, *305*, 787–796.
22. Gorban, A.N.; Popova, T.G.; Sadovsky, M.G. Classification of Symbol Sequences over Their Frequency Dictionaries: Towards the Connection between Structure and Natural Taxonomy. *Open Systems & Information Dynamics* **2000**, *7*, 1–17. doi:10.1023/A:1009652616706.
23. Sadovsky, M.G. Comparison of Real Frequencies of Strings vs. the Expected Ones Reveals the Information Capacity of Macromoleculae. *Journal of Biological Physics* **2003**, *29*, 23–38. doi:10.1023/A:1022554613105.
24. Sadovsky, M.G.; Putintseva, J.A.; Shchepanovsky, A.S. Genes, information and sense: complexity and knowledge retrieval. *Theory in Biosciences* **2008**, *127*, 69–78. doi:10.1007/s12064-008-0032-1.

25. Kohany, O.; Gentles, A.J.; Hankus, L.; Jurka, J. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC bioinformatics* **2006**, *7*, 1–7.
26. Salemme, M.; Sica, M.; Gaudio, L.; Aceto, S. The OitaAG and OitaSTK genes of the orchid Orchis italica: a comparative analysis with other C-and D-class MADS-box genes. *Molecular biology reports* **2013**, *40*, 3523–3535.
27. Hannat, S.; Pontarotti, P.; Colson, P.; Kuhn, M.L.; Galiana, E.; La Scola, B.; Aherfi, S.; Panabières, F. Diverse trajectories drive the expression of a giant virus in the oomycete plant pathogen Phytophthora parasitica. *Frontiers in microbiology* **2021**, *12*.
28. Gorban, A.N.; Zinovyev, A.Y. Principal Manifolds for Data Visualisation and Dimension Reduction. In *Lecture Notes in Computational Science and Engineering*, 2nd ed.; Gorban, A.N.; Kégl, B.; Wünsch, D.; Zinovyev, A.Y., Eds.; Springer: Berlin – Heidelberg – New York, 2007; Vol. 58, pp. 153–176.
29. Gorban, A.N.; Zinovyev, A.Y. Fast and user-friendly non-linear principal manifold learning by method of elastic maps. 2015 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2015, Campus des Cordeliers, Paris, France, October 19-21, 2015, 2015, pp. 1–9.
30. Miller, F.; Vandome, A.; McBrewster, J. *Levenshtein Distance*; VDM Publishing, 2009.
31. Chakraborty, D.; Goldenberg, E.; Koucký, M. Streaming algorithms for embedding and computing edit distance in the low distance regime. Proceedings of the forty-eighth annual ACM symposium on Theory of Computing. ACM, 2016, pp. 712–725.
32. Kaur, Y.; Sohi, N. Comparison of Different Sequence Alignment Methods – A Survey. *International Journal of Advanced Research in Computer Science* **2017**, *8*.
33. Wang, X.D.; Liu, J.X.; Xu, Y.; Zhang, J. A survey of multiple sequence alignment techniques. International Conference on Intelligent Computing. Springer, 2015, pp. 529–538.
34. Sadovsky, M.G.; Senashova, M.Y.; Malyshev, A.V. Amazing symmetrical clustering in chloroplast genomes. *BMC bioinformatics* **2020**, *21*, 1–14.
35. Sadovsky, M.G.; Putintseva, Y.A.; Senashova, M.Y. Eight clusters, synchrony of evolution and unique symmetry in chloroplast genomes: The offering from triplets. *CHLOROPLASTS AND CYTOPLASM* **2018**, p. 25.
36. Sadovsky, M.; Putintseva, Y.; Chernyshova, A.; Fedotova, V. Genome Structure of Organelles Strongly Relates to Taxonomy of Bearers. Bioinformatics and Biomedical Engineering; Ortuño, F.; Rojas, I., Eds.; Springer International Publishing: Cham, 2015; pp. 481–490.
37. Sadovsky, M.; Putintseva, Y.; Birukov, V.; Novikova, S.; Krutovsky, K. *De Novo* Assembly and Cluster Analysis of Siberian Larch Transcriptome and Genome. Bioinformatics and Biomedical Engineering; Ortuño, F.; Rojas, I., Eds.; Springer International Publishing: Cham, 2016; pp. 455–464.
38. Gorban, A.N.; Popova, T.G.; Zinovyev, A.Y. Seven clusters in genomic triplet distributions. *In Silico Biology* **2003**, *3*, 471–482.
39. Gorban, A.N.; Popova, T.G.; Zinovyev, A.Y. Four basic symmetry types in the universal 7-cluster structure of microbial genomic sequences. *In Silico Biology* **2005**, *5*, 265–282.
40. Gorban, A.; Popova, T.; Zinovyev, A. Codon usage trajectories and 7-cluster structure of 143 complete bacterial genomic sequences. *Physica A: Statistical Mechanics and its Applications* **2005**, *353*, 365 – 387. doi:https://doi.org/10.1016/j.physa.2005.01.043.
41. Fedotovskaya, V.; Sadovsky, M.; Kolesnikova, A.; Shpagina, T.; Putintseva, Y. Function vs. Taxonomy: Further Reading from Fungal Mitochondrial ATP Synthases. IWBBIO, 2020, pp. 438–444.
42. Sadovsky, M.; Fedotovskaya, V.; Kolesnikova, A.; Shpagina, T.; Putintseva, Y. Function vs. taxonomy: the case of fungi mitochondria ATP synthase genes. International Work-Conference on Bioinformatics and Biomedical Engineering. Springer, 2019, pp. 335–345.
43. Sadovsky, M.G.; Senashova, M.Y.; Gorban, I.K.; Gustov, V.S. Non-Coding Regions of Chloroplast Genomes Exhibit a Structuredness of Five Types. International Work-Conference on Bioinformatics and Biomedical Engineering. Springer, 2019, pp. 346–355.