Genome Structure of Organelles Strongly Relates to Taxonomy of Bearers

Michael Sadovsky, Yulia Putintseva, Anna Chernyshova, and Vaselina Fedotova

Institute of Computational Modelling of SB RAS,
Akademgorodok, 660036 Krasnoyarsk, Russia
msad@icm.krasn.ru, kinomanka85@mail.ru,
{anna12651,vasilinyushechka}@gmail.com
http://icm.krasn.ru

Abstract. We studied the relations between the triplet frequency dictionaries of organelle genome, and the phylogeny of their bearers. The clusters in 63-dimensional space were identified through K-means, and the clade composition of those clusters has been investigated. Very high regularity in genomes distribution among the clusters was found, in terms of taxonomy. The strong synchrony in evolution of nuclear and organelle genomes manifests through this correlation: the proximity in frequency space was determined over the organelle genomes, while the proximity in taxonomy was determined morphologically. Similar effect is also found in the ensembles of other (say, yeast) genomes.

Keywords: frequency, triplet, order, cluster, similitude, elastic map, morphology, evolution, synchrony.

1 Introduction

Statistical properties of nucleotide sequences may tell a lot to a researcher. The patterns observed in sequences correlate to functions encoded in a sequence, or to a taxonomy of a bearer of that latter. Here we shall study those correlations between the structure, and the taxonomy.

A variety of patterns in a nucleotide sequence is tremendous. Here a consistent and comprehensive study of frequency dictionaries answers some questions concerning the statistical and information properties of DNA sequences. A frequency dictionary, whatever one understands for it, is rather multidimensional entity. That latter is supposed to be the simplest structure. Further, we shall concentrate on the study of the frequency dictionaries [8–10] of the thickness $q = 3$; in other words, the triplet composition only will be taken into consideration. Here we studied the *structure – taxonomy* relations for mitochondrion vs. host genomes, and chloroplast vs. host genomes.

Let now introduce more strict definitions and issues. Consider a continuous symbol sequence of the length N (total number of symbols in it) from four-letter alphabet $\aleph = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$; such sequence represents some genetic entity (genome, chromosome, etc.). We stipulate that no other symbols or gaps in the

F. Ortuño and I. Rojas (Eds.): IWBBIO 2015, Part I, LNCS 9043, pp. 481–490, 2015.

sequence take place. Any coherent string $\omega = \nu_1\nu_2 \ldots \nu_q$ of the length q makes a word. A set of all the words occurred within a sequence yields the support of that latter. Counting the numbers of copies n_ω of the words, one gets a finite dictionary; changing the numbers for the frequency

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

one gets the frequency dictionary W_q of the thickness q. This is the main object of our study.

Further, we shall concentrate on frequency dictionaries W_3 (i. e., the triplet composition) only. Thus, a frequency dictionary W_3 calculation converts any genetic entity into a point in (formally) 64-dimensional metric space. Obviously, two genetic entities with identical frequency dictionaries $W_3^{(1)}$ and $W_3^{(2)}$ are mapped into the same point in the space. On the contrary, the absolute congruency of two frequency dictionaries $W_3^{(1)} = W_3^{(2)}$ does not guarantee a complete coincidence of the original sequences. Nonetheless, such two sequences are indistinguishable from the point of view of their triplet composition.

Definitely, few entities may have very proximal frequencies of all the triplets, but few others may have not, thus making a distribution of the points in 64-dimensional space inhomogeneous. So, the key question here is what is the pattern of this distribution of mitochondrion genomes in that space? Are there some discrete clusters, and if yes is there a correlation to a phylogeny of the genome bearers and clusters? Some results preliminary answering this question could be found in [8, 5, 6].

To address the questions, we have implemented an unsupervised classification of both mitochondrion and chloroplast genomes, in (metric) space of frequencies of triplets. There were implemented a series of clusterizations, for two, three, four, ..., eight clusters, for both types of genomes. Then, the taxa composition of the classes has been studied; moreover, the relation between the clusters was specially studied, when we changed a clusterization in K clusters for that one in $K - 1$ clusters. Besides, a considerable correlation in taxa composition was found, for the observed clusterizations. Briefly speaking, these correlations prove the high synchrony in the evolution of two (physically) independent genetic systems: somatic one, and the organelle one.

This paper presents the evidences of the strong synchrony in evolution of mitochondrion genomes and nuclear ones, as well as the synchrony in evolution of chloroplast genomes vs. nuclear ones.

2 Material and Methods

2.1 Genetic Sequences

All the sequences were retrieved from EMBL–bank. **"Junk" symbol agreement**: some entries contain the "junk" symbols (those that fall beyond the original alphabet $\aleph$). All such symbols have been omitted furthered with the concatination of the obtained fragment into an entity.

Originally, the release used to retrieve mitochondria genomes containes $\sim 6.4 \times 10^3$ entries. The final database used in our study enlists 3721 entries. Similarly, the full list of chloroplast genomes at the release used for them exceeded five hundred entries; the study on chloroplast clusterization has been carried out with 251 genomes. This discrimination comes from the (not obvious) constraint: we had to eliminate from the study the entries which represent rather highly ranked clade solely, as the single species in the clade. A highly order clade presented with a single genome (that is a single species) yields a "signal" strong enough to deteriorate a general pattern, but weak one to produce distinguishable details in the distribution pattern. Thus, we enlist into the final databases the entries representing an order (and higher clades) with five species or more, for mitochondria. Similar cut-off number for chloroplast genomes was equal to 3 species.

Table 1. Mitochondria database structure; M is the abundance of the clade

Order	M	Order	M	Order	M	Order	M
Actinopterygii	1181	*Amphibia*	151	*Anthozoa*	16	*Arachnida*	10
Aves	197	*Bivalvia*	34	*Cephalopoda*	45	*Cestoda*	28
Chromadorea	5	*Gastropoda*	16	*Homoscleromorpha*	14	*Insecta*	350
Malacostraca	33	*Mammalia*	1457	*Reptilia*	172	*Trematoda*	13

The stricture of the final mitochondrion database is shown in Table 1. Since the total number of chloroplast genomes under consideration is significantly less, in comparison to the mitochondria list, the clade composition of that former seems to be less apparent; meanwhile, the chloroplast database includes 157 broadleaf species against 94 conifer ones.

2.2 Clusterization Methods

We implemented unsupervised classification by K-means to develop classes (see details and a lot of examples in [3, 2, 11, 7]). To cluster, we had to reduce the data space dimension to 63: the reduction results from the equality to 1 of the sum of all frequencies. Formally speaking, any triplet could be excluded from the data set; practically, we excluded the triplets (specific, for mitochondria, and for chloroplasts) that yield the least standard deviation, over the set of genomes under consideration. Evidently, such triplets make least contribution into the distinguishing the entities, in the space of frequencies.

A K-means implementation may be based on a number of distances; here we used Euclidean distance. Also, no class separability has been checked. All the results were obtained with *ViDaExpert* software by A. Zinovyev[1].

[1] `http://bioinfo-out.curie.fr/projects/vidaexpert/`

"Downward" vs. "Upward" Classification. Two versions of the K-means classification implementation could be developed: "downward" vs. "upward" ones, respectively. They both are based on a standard K-means technique, but the difference is in the mutual interaction between the clusterizations developed for different number of clusters.

"Downward" classification. This kind of classification is designed to follow the classical morphology based classification. It starts from the clusterization of the entire set of genomes (frequency dictionaries) into the minimal number $\mathfrak{M}_c$ of clusters with the given stability of the clusterization. That latter is understood as the given number of volatile genomes, i. e. genomes that may change their cluster attribution with any new clustering realization. Then each of the clusters is to be separated into the similar (i. e. minimal stable subclusters) set of subclusters, etc. The procedure is to be trunked at the given "depth" of the cluster separation, usually determined by the volatility of a significant part of genomes.

Table 2. Standard deviation figures, for mitochondria and chloroplast databases

chloroplasts				mitochondria			
GAC	0,000540	ACC	0,000672	GCG	0,001329	TCG	0,001712
GGC	0,000593	GTC	0,000731	CGT	0,001608	GTC	0,001715
GCC	0,000612	CGC	0,000748	CGA	0,001690	AGG	0,001726

Thus, a "downward" classification yields the structure that is a tree, so making it close to a standard morphological classification.

"Upward" classification. On the contrary, the upward classification consists in the separation of the entire set of genomes, sequentially, into the series of clusters

$$\mathfrak{C}_2, \mathfrak{C}_3, \mathfrak{C}_4, \ldots, \ \mathfrak{C}_{K-1}, \mathfrak{C}_K \, .$$

Here we assume that the clusterization $\mathfrak{C}_2$ is stable. Again, the series is to be trunked at the given number K; we put $K = 8$ in this study with chloroplasts.

The key question here is the mutual relation between the members of a cluster $\mathfrak{C}(l)_j$ from $\{\mathfrak{C}(i)_j\}$ $(1 \leqslant i \leqslant j)$ clusterization with the clusters from $\{\mathfrak{C}(m)_{j-1}\}$ $(1 \leqslant m \leqslant j-1)$ clusterization. Here the index l enlists the clusters at the $\{\mathfrak{C}(i)_j\}$ clusterization. There could be (roughly) three options:

- A cluster $\mathfrak{C}(n)_j$ is entirely embedded into the cluster $\mathfrak{C}(l)_{j-1}$, with some l and j;
- The greater part of the members of a cluster $\mathfrak{C}(n)_j$ is embedded into the cluster $\mathfrak{C}(l)_{j-1}$, but the minor part is embedded into the other cluster $\mathfrak{C}(m)_{j-1}$;
- A cluster $\mathfrak{C}(n)_j$ is almost randomly spread between the set of clusters $\mathfrak{C}(l)_{j-1}$, $l = 1, 2, \ldots, l^*$.

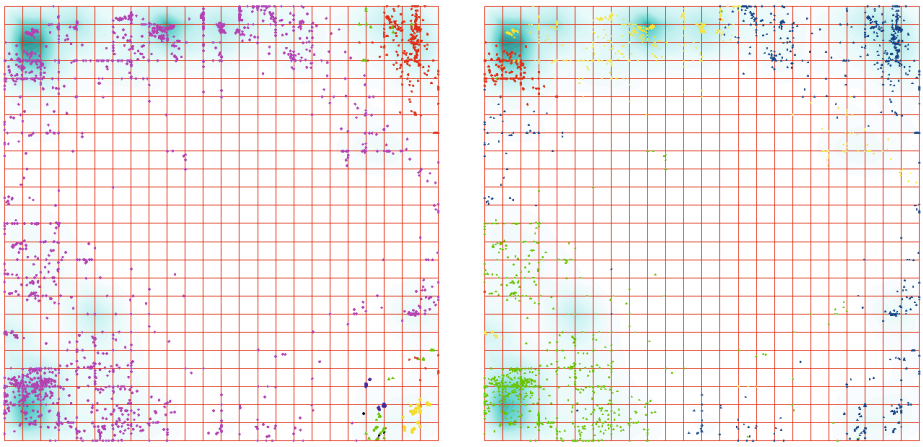

Fig. 1. Soft 25 × 25 elastic map for 3954 mitochondria genomes; left is the types distribution, right is the clades distribution for chordata type. See the text for details.

Thus, an upward classification yields a pattern that is a graph with cycles; at the worst case, the graph is fully connected, and here no essential structuredness is observed. If the graph has rather small number of cycles, then it reveals the relations between the clusters (determined through the proximity in frequency space), and the taxonomy (determined over the nuclear genome).

2.3 On the Stability of *K*-means Clusterization

Another essential point is the volatility of genomes (in K-means clusterization): that is equivalent to the clusterization stability. Speaking on stability, we stipulate some genomes always (or almost always) occupy the same cluster, for different starting distributions. Other genomes tend to change their cluster position, in a series of K-means implementations.

Thus, the former set of entities is supposed to be stable, while the latter gathers the unstable entities. Stability here could be evaluated through the portion of genomes always occupying the cluster together; volatile genomes, on the contrary, change their cluster occupation, for different implementations of K-means. Everywhere further we shall stipulate that stability in K-means clusterization means that the stable ensemble of genomes exhibits the same clusterization pattern in 0.85-part of the series of K-means implementations.

3 Results

We studied the relation between the structure defined in terms of a triplet composition of organelle genomes, and the taxonomy determined according to a morphology, for two types of organelles: chloroplasts and mitochondria. Everywhere

below the upward classifications only are considered, both for mitochondria and chloroplasts. No class separation conditions has been checked, in both genomes databases.

Besides, the number of classes $\mathfrak{C}(n)$ varied form two to eight: $2 \leqslant \mathfrak{C}_K \leqslant 8$, for both databases. All these constraints have been put on mostly due to technical reasons.

3.1 Mitochondria

Mitochondria genomes database, unlike the chloroplasts one, is quite abundant; on the other hand, it is rather biased: Table 1 shows this fact. Some genera are overrepresented, in comparison to others, but others seem to be underrepresented. Such bias affects the K-means clusterization. In particular, it may result in a stability decay of the clusterization developed by K-means technique. Due to this discrepancy, we have used elastic map technique to figure out the clusters in triplet frequency space.

Elastic map is another powerful approach to visualize and analyze multidimensional data. This approach makes no way to establish either upward, or downward classification: it yields a distribution of genomes on a non-linear two-dimensional manifold (elastic map). We have used the detailed soft map of 25×25 size. Figure 1 shows the distribution of the entities over the map; left part of the figure presents the distribution of *Cordata* (the most abundant class) in pink ring labels, *Arthropoda* in red squares, *Mollusca* in green triangles, *Nematoda* are shown in brick-like colored diamonds, flat warms are whosn in sand-colored pentagons and finally *Porifera* are shown in dark-blue hexagons.

The right part of the figure shows the distribution of three main clades of *Chordata* type: *Mammalia* are shown in yellow diamonds, birds are shown in red pentagons, and fishes are shown in green triangles.

Color background indicates the average local density of the genomes in this map. One may see quite unexpected growth of the local density; that former is located at the map node $[5, 8]$, if the lowest left one is supposed to be the $[1, 1]$ node.

3.2 Chloroplasts

Figure 3 shows the graph of embedments for the clusters, where the number of these latter changed from 8 to 3. We developed the clusterizations for three, four, ..., eight clusters, and studied the composition of each cluster, at the each "depth". The key question was whether the species tend to keep together, when the number of clusters in a clusterization is decreased.

The Figure answers distinctly and apparently this question: the clades in the boxes correspond to genera, while the species (not shown in the figure) always make a solid group, when changing the number of clusters. Thus, boxes having two upright arrows showing the transfer of entities from $\mathfrak{C}_l$ clusterization to $\mathfrak{C}_{l-1}$ one contain two groups of species belonging the same genus (or family).

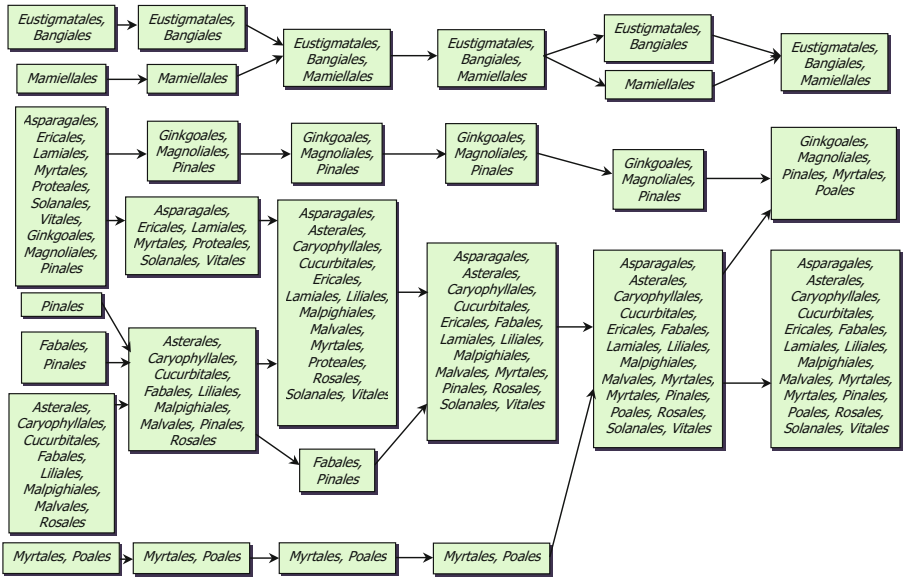

Fig. 2. The chain of clusters identified through K-means, with high stability level. Only orders are shown in the graph.

The point is that the scheme shown in Fig. 3 is just a part of a general pattern: the figure shows the genomes that exhibit reasonable stability level in the upward clusterization. There a sounding number of chloroplast genomes that exhibit a low stability, for various K-means clusterizations. It should be noticed that the ensemble of unstable genomes may change, as the number of clusters goes down from 8 to 3. Figure 2 shows this part of genomes. Careful examination of Fig. 2 has the following formula for cluster composition: 7, 6, 5, 4, 4, 3 clusters stably identified over the dataset. This formula comes from the degeneration of a stable cluster, when the clusterization over 8, 7, 6 and 5 clusters is carried out (on the contrary to Fig.3). In other words, an attempt to create an 8-class K-means clusterization yields seven stable clusters while the eighth one is opportunistic: it comprises various genomes that tend to occupy the different clusters, for different realizations of K-means.

The graph shown in Fig. 2 is not connected: the branch comprising the algae (both green and red ones) is isolated, at any depth of a classification (orders *Eustigmatales*, *Bangiales* and *Mamiellalis*.) Also, one can see that some orders occupy two clusters, at the depth 8 and some others. It means that this order is split: some species occupy other cluster than others.

An implementation of the clusterization through K-means for two classes yields the distinct and clear separation of algae from all other plants; the stability of such clusterization is not too high. A series of a thousand realizations of K-means with two-class separation exhibits about 680 realizations with discrete

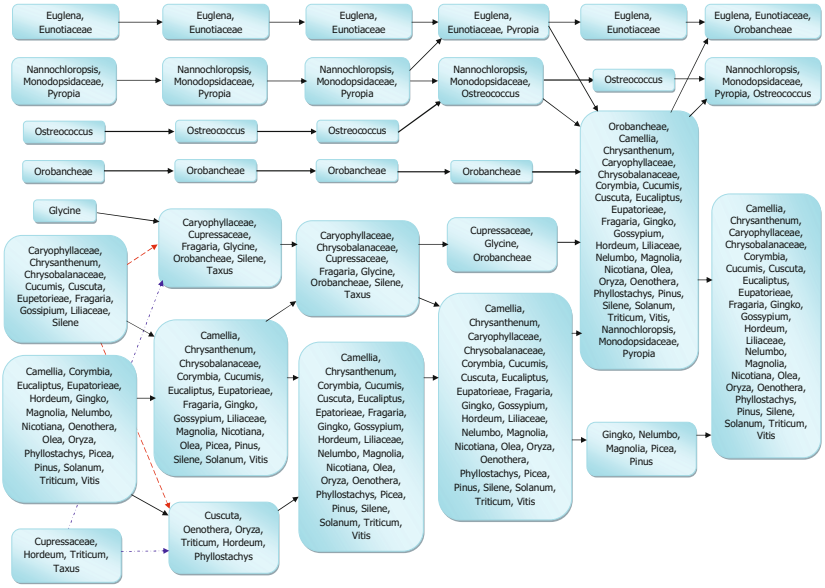

Fig. 3. Pattern of the "upward" embedment of various clades of plants in the unstable classification, developed over chloroplast genomes

isolation of algae from the other plants, while other part of realizations may combine some algae with higher plants.

Stable ensemble of chloroplast genomes differs from the unstable one in the relevant graph connectivity. Stable genomes exhibit significantly less complexity of the pattern: there is an isolated subgraph making the entire graph disconnected. Moreover, this subgraph have very simple structure (linear or almost linear) and is disjointed from the main body of that former. On the contrary, the graph representing the unstable genomes is connected: there are no isolated subgraphs or any other parts.

4 Discussion

To begin with, consider Table 2 in detail. Evidently, the figures for the standard deviation observed for mitochondria exceed the similar figures for chloroplasts in order. Nonetheless, this difference remains the same, for any subbase to be developed from the original one. Probably, such significant difference in the figures results from the fact that mitochondria exhibit the highest possible violation of Chargaff's parity rule, among all other genetic entities.

A study presented in this paper is done within the scope of population genomics methodology. The most intriguing result of the study is the very high correlation between the statistically identified clusters of genomes, and their

taxonomy reference. The key point is that we used organelle genomes to derive the clusterization, while the taxonomy was determined traditionally, through morphology, which is ultimately defined by a nuclear genomes. There is no immediate interaction between the nuclear and organelle genomes. The study has been carried out for both main organelles: chloroplasts and mitochondria.

All mitochondria have the same function; same is true for chloroplasts. Thus, the impact of a function divergence was eliminated in our studies. Probably, a database structure is crucial in this kind of studies. We have used an unsupervised classification technique to develop a distribution of genomes into few groups. The results of such classification are usually quite sensitive to an original database composition [7]. Luckily, the genetic banks are rapidly enriched with newly deciphered genomes of organelles, so the stable and comprehensive results showing the reliable relation between structure and taxonomy could be obtained pretty soon. Moreover, a growth of genetic database may provide a comprehensive implementation of a "downward" classification.

The approach presented above looks very fruitful and powerful. One can expand the approach for the following problems to be solved:

- To study the clusterizations as described above for the database consisting of the genomes of mitochondria, and chloroplasts, of the same species. This study would unambiguously address the question on the relation between structure and function: since the organelle genomes under consideration would belong the same organisms, one may expect an elimination of the taxonomy impact, on the results. Meanwhile, this point should be carefully checked, since the results might be sensitive to the list of species involved into the study;
- To study the clusterization of the frequency dictionaries corresponding to the individual genes (or genes combinations) retrieved from the raw genomes of organelles. The clusterization of such genetic entities would address the question on the mutual interaction in a triadic pattern *structure – function – taxonomy*.

References

1. Bugaenko, N.N., Gorban, A.N., Sadovsky, M.G.: Open Systems & Information Dyn. 5, 265–281 (1998)
2. Gorban, A.N., Zinovyev, A.Y.: Int. J. of Neural Systems 20, 219 (2010)
3. Gorban, A.N., Kegl, B., Wünsch, D.C., Zinovyev A.Y. (eds.) Principal Manifolds for Data Visualisation and Dimension Reduction. LNCSE, vol. 58, p. 332. Springer, Heidelberg (2007)
4. Gorban, A.N., Popova, T.G., Zinovyev, A.Y.: In Silico Biology 3, 471 (2003)
5. Gorban, A.N., Popova, T.G., Sadovsky, M.G., Wünsch, D.C.: Information content of the frequency dictionaries, reconstruction, transformation and classification of dictionaries and genetic texts. In: Intelligent Engineering Systems through Artificial Neural Networks, 11 — Smart Engineering System Design, pp. 657–663. ASME Press, New York (2001)

6. Gorban, A.N., Popova, T.G., Sadovsky, M.G.: Open Systems & Information Dyn. 7, 1 (2000)
7. Fukunaga, K.: Introduction to statistical pattern recognition. 2nd edn., 591 p. Academic Press, London (1990)
8. Sadovsky, M.G., Shchepanovsky, A.S., Putintzeva Yu, A.: Theory in Biosciences 127, 69 (2008)
9. Sadovsky, M.G.: J. of Biol. Physics 29, 23 (2003)
10. Sadovsky, M.G.: Bulletin of Math. Biology 68, 156 (2006)
11. Shi, Y., Gorban, A.N., Yang, T.Y.: J. Phys. Conf. Ser. 490, 012082 (2014)