Intrinsic Structuredness of Mitochondria Genomes*

Roman Kosarev[1][0000–0002–0885–599X], Maria Senashova[2][0000–0002–1023–7103] and Mikhail Sadovsky[2][0000–0002–1807–0715]

[1] Siberian Federal University, Krasnoyarsk 660041, Russia
[2] Institute of Computational Modeling SB RAS, Krasnoyarsk 660036, Russia
{msen,msad}@icm.krasn.ru

Abstract. Previously, a seven-cluster pattern for bacterial genomes has been reported. The pattern is revealed through the distribution of the formally identified short fragments of a genome converted into triplet frequency dictionaries. Later, similar patterns have been observed for chloroplast genomes, with their comparison with the patterns observed for cyanobacteria genomes revealing the difference in the symmetry of the patterns. Here, we apply the same methodology to reveal a pattern for mitochondria genomes. Six types of patterns have been found. Some specific violations in the symmetry of the patterns is discussed.

Keywords: Order, Distribution, Clustering, Evolution, Symmetry.

1 Introduction

A diversity and role of various structures to be found in biological macromolecules are still challenging for researchers working in the areas ranging from classic biology to mathematics and computer science. One may follow two different paradigms in such a study: the former is a structure – function interplay, and the latter is the evolutionary exploration. Revealing the details of interaction between the structure of the DNA sequences and encoded functions is the key issue of modern system biology which is far from being solved. Moreover, researchers find new structure patterns, or reveal a new interplay between the known patterns and functions; and tremendous growth in the relevant techniques makes the problem even more profound.

The importance of the evolutionary insight of such studies is also beyond any doubt. The study of the interplay between the structure of the DNA molecule and the encoded functions brings new knowledge on the dynamics and/or evolutionary processes occurring in various biological systems ranging from a cell to communities.

A choice of the biological matter for such studies may raise a problem. Without mentioning the sequencing, assembling, annotation etc. errors, one faces an extremely high complexity of the objects under consideration. Investigating molecular biological

and genetic entities, one has to deal with a triade of issues which are *structure*, *function* and *phylogeny* (in a wider sense). The issues exhibit a strong interplay and heavily influence each other. It should be emphasized that it is often impossible to distinguish them and figure out the independent impact of a particular issue.

From this viewpoint, prokaryotic organisms seem to be more convenient for the studies. Their genomes are significantly shorter in comparison to eukaryotic ones and (almost always) consist of a single circular chromosome. Organelle genomes are even more advantageous when compared to prokaryotic ones; they encode the same function. So, no function impact on the interplay mentioned above is expected, if a researcher studies the organelle genomes of the same group.

A number of papers [16, 2, 4, 3, 1, 7, 15, 14] addresses the problems mentioned above. Meanwhile, it should be emphasized that these studies mainly consider the role of functions encoded in the genomes. We focus on the structure of mitochondrial genomes understood as a pattern provided by clustering of the formally idenfied fragments of a genome. Previously, such studies were carried out on bacteria [6, 5], transcriptome [9, 11, 12] and chloroplasts [13, 8, 10]. Here, we concentrate on mitochondrial genomes, which have not been investigated yet, from the viewpoint of the inner structuredness provided by the comparison of the statistical properties of the formally identified fragments.

2 Material and methods

2.1 Tiling

Consider a genetic sequence of the length L from the four-letter alphabet $\aleph = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$. No other symbols are stipulated to occur in the sequence. To reveal the structuredness, tiling is developed. This is a set of (overlapping) fragments of the given length Δ identified within the sequence with the move step t; we take $\Delta = 603$ and $t = 11$. Next, each tile is converted into the triplet frequency dictionary $W_{(3,3)}(j)$; here, j enumerates the tile location along the sequence.

The triplet frequency dictionary $W_{(3,3)}$ is a list of all the triplets $\omega = \nu_1\nu_2\nu_3$ ranging from $\omega = \mathrm{AAA}$ to $\omega = \mathrm{TTT}$ supplied with their frequency figure f_ω. The triplets are counted along the fragment, with the reading frame shift equal to 3. In other words, the triplets are counted with neither gaps, nor overlaps. Further, we shall omit the subscript in the dictionary notation unless it makes a confusion. As soon as each fragment is converted into the frequency dictionary, the sequence is transformed into a set of points in the 64-dimensional metric space.

For the purposes of the study, each point was labeled with the location coordinate which is the number of the central nucleotide of the relevant tile in the sequence, and the relative phase index. The latter represents the location of the tile against the coding and non-coding regions found in the genome. To begin with, we neglected the exon-intron structure of the genes, and consider them as a solid coding region. To identify it we followed the annotation of the genome.

There are seven labels of the index: J, F_0, F_1, F_2, B_0, B_1 and B_2. The tile index is J, if its central nucleotide falls out of a coding region. The tile is indexed as F_k, $0 \le k \le 2$, if the central nucleotide falls inside the coding region and the distance from the starting nucleotide of the coding region to the central one yields the remainder equal to k when divided by 3. This labeling holds for the genes located in the leading strand. Similarly, the tile is labeled with B_k, $0 \le k \le 2$, if the gene is located at the ladder strand; the distance here is determined from the end of the coding region, and it is counted in an opposite direction, since the genetic sequence is presented in the genetic bank with the leading strand only.

2.2 Triplet Exclusion

A triplet frequency dictionary maps a tile into a point in the 64-dimensional metric space. The problem is that the sum of all the frequencies is one:

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

making the frequencies linearly dependent. This linear constraint may cause a false signal when clustering, so a triplet must be excluded from the analysis. Formally, any triplet may be excluded; in fact, we excluded the triplet yielding the least standard deviation determined over the set of tiles for each genome.

2.3 Genetic Data

The genomes for the study were downloaded from the EMBL–bank; Table 1 shows the clade distribution over the genetic matter.

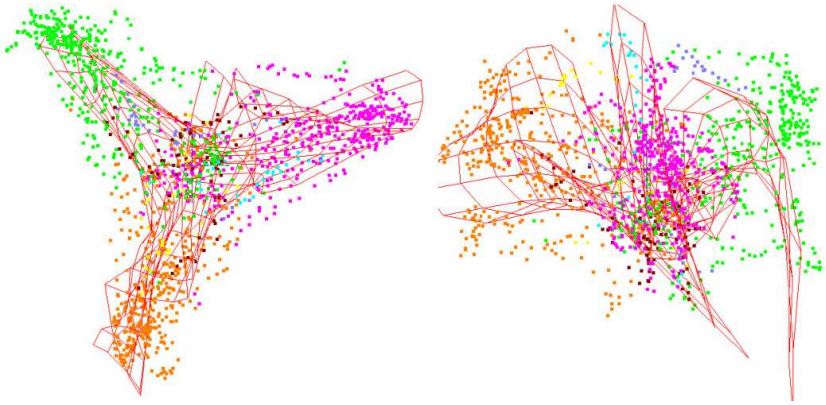
Table 1. Clade composition of the studied genomes. SC is a systematic category, SG is a systematic group, M is the group abundance.

SC	SG	*M*	SC	SG	*M*
type	Fungi	24	type	Mollusca	25
class	Cyclostomata	14	class	Chondrichthyes	25
type	Platyhelminthes	25	type	Roundworms	25
class	Crustacea	25	class	Arachnida	25
class	Reptilia	24	class	Birds	25
kingdom	Plants	35	kingdom	Higher fungi	25
division	Lichens	13	class	Amphybia	25
class	Osteichthyes	25	type	Annelida	25
class	Insects	25	class	Mammalia	25
division	Mosses	25			

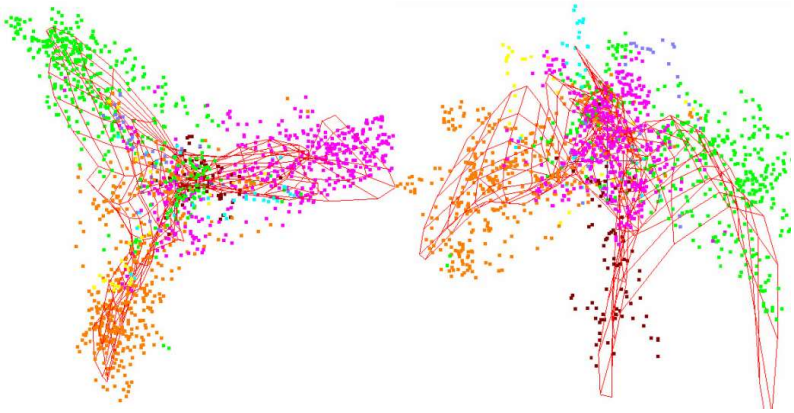
2.4 Clustering and Visualization

We used the freely distributed software *VidaExpert*[1] to visualize the patterns in the metric space. To do this, we used the Euclidean metrics. We studied the distribution of the tiling points in the principal component space. Everywhere below, the following coloring label system for the relative phase indices is applied:

- J phase tiles (corresponding to the non-coding regions) are colored in brown;
- the tiles indexed as F_0 and B_0 are colored in rose and violet, correspondingly;
- the tiles indexed as F_1 and B_1 are colored in green and cyan, correspondingly;
- and finally, the tiles indexed as F_2 and B_2 are colored in orange and yellow, correspondingly.

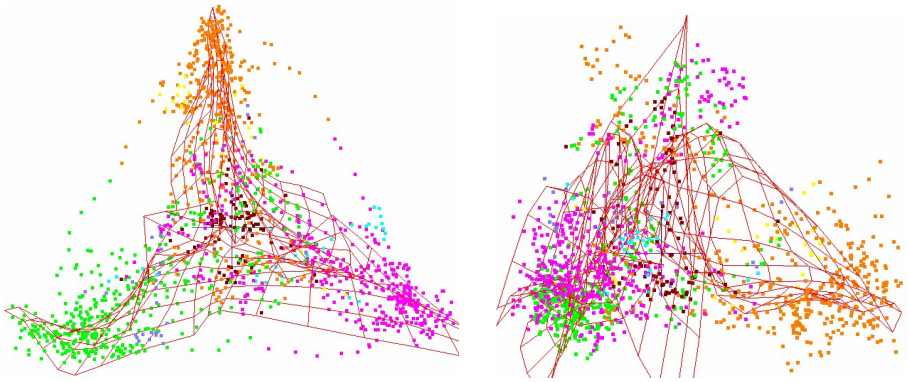


a)



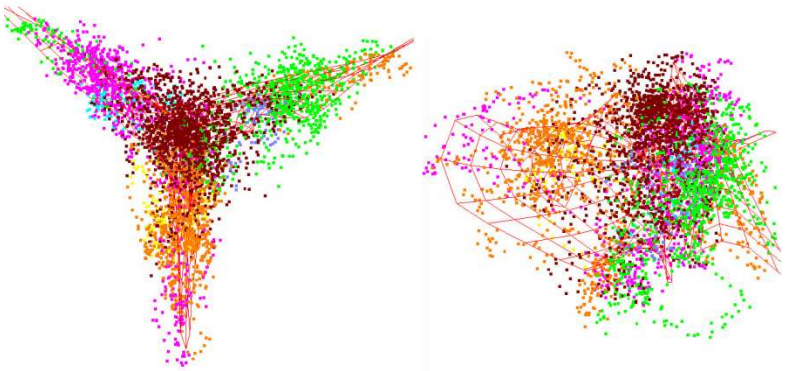
b)

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

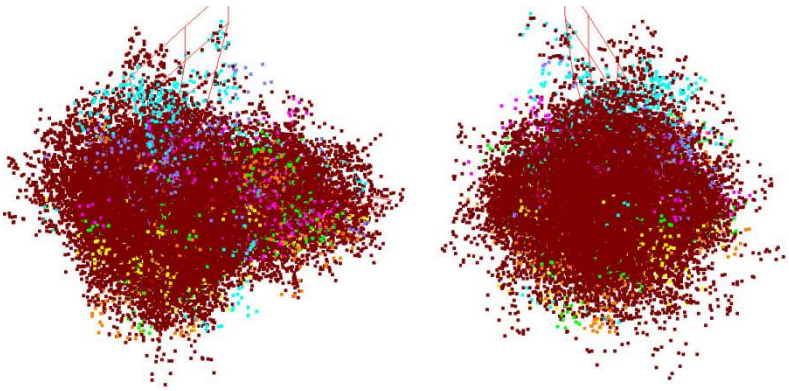


c)

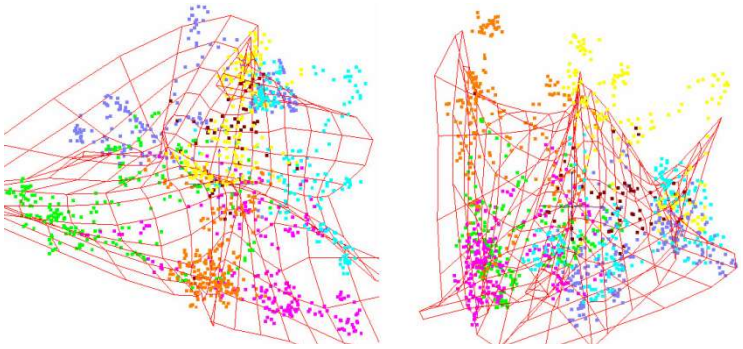
Fig. 1. Three types of the patterns observed in the mitochondrial genomes.



a)



b)



c)

Fig. 2. Three other types of the patterns observed in mitochondrial genomes.

Table 2. Distribution of the pattern types over the clades, absolute figures. Percentage is shown in parentheses. Types of patterns: I – three-beams, straight; II – three-beams, tailed; III – three-beams with nimbus; IV – three-beams with nucleus; V – ball; VI – amorphous.

Class	I	II	III	IV	V	VI
Amphybia	24 (96 %)	1(4 %)				
Osteichthyes	21 (84 %)		4 (16 %)			
Cyclostomata	12 (86 %)	2 (14 %)				
Mammalia	18 (76 %)	5 (20 %)				1 (4 %)
Birds	15 (64 %)	9 (36 %)				
Reptilia	13 (52 %)	10 (38 %)				
Chondrichthyes	15 (60 %)	2 (8 %)	8 (32 %)			
Fungi	23 (92 %)	2 (8 %)				
Annelida	23 (92 %)	1 (4 %)	1 (4 %)			
Roundworms	18 (72 %)	5 (20 %)				2 (8 %)
Mollusca	19 (76 %)	1 (4 %)	2 (8 %)			3 (12 %)
Platyhelminthes	21 (84 %)	3 (12 %)				1 (4 %)
Insects	19 (76 %)		6 (24 %)			
Arachnida	18 (72 %)		2 (8 %)			5 (20 %)
Crustacea	19 (76 %)		5 (20 %)			1 (4 %)
Plants	12 (30 %)	3 (7 %)			19 (46 %)	7 (17 %)
Higher fungi	9 (36 %)			15 (60 %)		1 (4 %)
Mosses				11 (44 %)	14 (56 %)	
Lichens	13 (100 %)					

3 Results and Discussion

For 19 clades, six types of the cluster patterns were found; Figs. 1 and 2 show that everywhere below in the Figs the left picture presents the projection in the (PC_1, PC_2) plane, and the right picture presents the projection in the (PC_2, PC_3) plane; here, PC

stands for the principal components. The following patterns are shown in Figs. 1 and 2:

- **Fig. 1(a)** shows the three-beam pattern of the mitochondrial genome of *Coelodonta antiquitatis*;
- **Fig. 1(b)** shows the three-beam tailed pattern of the mitochondrial genome of *Canis lupus*;
- **Fig. 1(c)** shows the three-beam pattern with nimbus for the mitochondrial genome of *Callorhinchus callorynchus*;
- **Fig. 2(a)** shows the three-beam pattern with nucleus for the mitochondrial genome of *Ganoderma lucidum*;
- **Fig. 2(b)** shows the ball-like pattern for the mitochondrial genome of *Oryza rufipogon*;
- **Fig. 2(c)** shows the amorphous structure, for the mitochondrial genome of *Heterometrus longimanus*.

The distribution of these six types of the above mentioned patterns over the clades is quite imbalanced: the former tends to occupy some specific clades rather than to spread homogeneously; Table 2 illustrates it. The table evidences that the straight three-beam pattern is peculiar for the majority of clades. Plants and mosses make an exception.

Two other patterns (these are three-beam tailed one and three-beams with nucleus) resemble to some extent those found in chloroplasts. Fungi are the leaders in the occurrence of three-beams with nucleus; the three-beam tailed pattern is also rather common in these clades, while insects and lower arthropodes, as well as mosses, lichens and higher fungi exhibit these patterns quite rarely.

Evidently, the chloroplast patterns differ from the mitochondrial ones in the number of fragments belonging to the leading and ladder strands, correspondingly. The mitochondria show a very low number of the latter. Platyhelminthes, higher fungi and lichens show the complete absense of such fragments. Besides, the chloroplasts differ from the mitochondria in the *tail* composition: the former comprises the fragments from the coding regions solely, while the latter comprises the fragments from the non-coding regions, exclusively.

Also, the mitochondria exhibit a formally new pattern called *three-beams with nimbus*. The nimbus consists of the fragments falling into the coding regions; the pattern looks like a thread (or several ones) surrounding the core and connecting the beams from the outside. The direct examination of the location of the fragments comprising the nimbus shows that they are successively located along the genome, as a rule. This structure is most common for fish and arthropodes.

Probably, the amorphous structure observed in some mitochiondrial genomes is the most amazing one. It has no symmetry observed over the other genomes, just making a structureless mess of points corresponding to the fragments. A remarkable fact is that this pattern is peculiar for plants and ancient mosses.

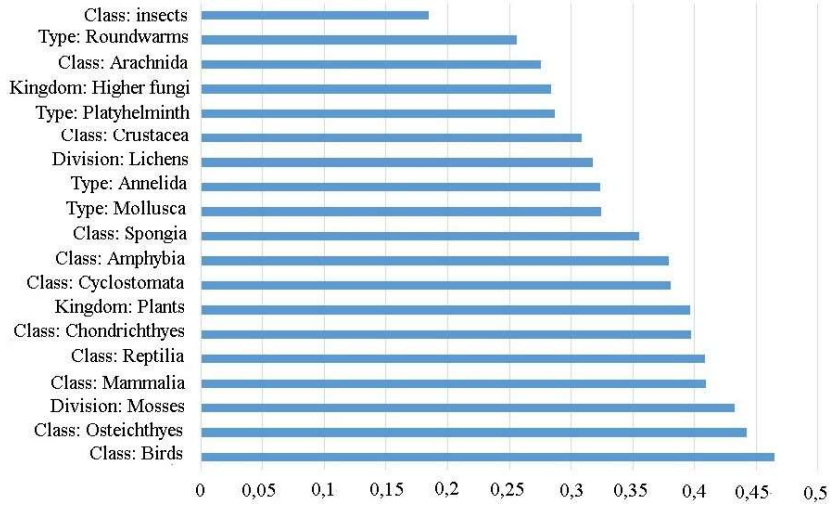


Fig. 3. Variation of the GC-content over the clades in descending order.

GC-content is claimed in [5, 6] to be the key factor determining the pattern type. It might be so for bacteria, but later it was shown that the GC-content has nothing to do with the type of pattern, for the chlorolplast genomes [10, 13]. Let us consider what is implied by the GC-content. It is a portion of nucleotides G and C if counted within a fragment of a genetic sequence. So, we traced the average GC-content over the clades to retrieve the interplay between the observed patterns and this value.

The highest GC-content was found for birds (0.46), to be followed by (in descending order) Osteichthyes, mosses, Mammalia and Reptilia. No evident and clear interplay was found in these data. Also, further progress may be achieved if a greater number of genomes is taken into consideration, especially with a kind of censorship in the taxa abundance observed in nature.

4 Conclusion

Both mitochondria and chloroplasts tend to possess the three-beam pattern (with variations); yet, no identity could be assumed between the chloroplasts and mitochondria, in spite of the total lack of the function diversity impact inside the group. Chloroplasts seem to be more rigorous genetic entities in terms of the pattern occurrence. On the contrary, mitochondrial genomes yield a wider diveristy of the patterns. This fact is likely to be due to the difference in the function of these two organelles and in the length of their genomes. Meanwhile, this problem requires further studies.

References

1. Adams, K.L., Palmer, J.D.: Evolution of mitochondrial gene content: gene loss and transfer to the nucleus. Molecular phylogenetics and evolution **29(3)**, 380–395 (2003)
2. Boore, J.L.: The use of genome-level characters for phylogenetic reconstruction. Trends in Ecology & Evolution **21(8)**, 439–446 (2006)
3. Chun, J., Rainey, F.A.: Integrating genomics into the taxonomy and systematics of the Bacteria and Archaea. International journal of systematic and evolutionary microbiology **64(2)**, 316–324 (2014)
4. Dagan, T.: Phylogenomic networks. Trends in microbiology **19(10)**, 483–491 (2011)
5. Gorban, A.N., Popova, T.G., Zinovyev, A.Y.: Seven clusters in genomic triplet distributions. In Silico Biology **3(4)**, 471–482 (2003). http://content.iospress.com/articles/in-silico-biology/isb00110
6. 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 **5**(3), 265–282 (2005). http://content.iospress.com/articles/in-silico-biology/isb00185
7. Koonin, E.V.: The origin and early evolution of eukaryotes in the light of phylogenomics. Genome biology **11(5)**, 209 (2010)
8. Sadovsky, M., Senashova, M., Malyshev, A.: Eight-cluster structure of chloroplast genomes differs from similar one observed for bacteria. ArXiv e-prints (Feb 2018)
9. Sadovsky, M., Putintseva, Y., Birukov, V., Novikova, S., Krutovsky, K.: De Novo assembly and cluster analysis of siberian larch transcriptome and genome. In: Ortuño, F., Rojas, I. (eds.) Bioinformatics and Biomedical Engineering. pp. 455–464. Springer International Publishing, Cham (2016)
10. Sadovsky, M., Senashova, M., Malyshev, A.: Chloroplast genomes exhibit eightcluster structuredness and mirror symmetry. In: Rojas, I., Ortu~no, F. (eds.) Bioinformatics and Biomedical Engineering. pp. 186-196. Springer International Publishing, Cham (2018)
11. Sadovsky, M.G., Birukov, V.V., Putintseva, Y.A., Oreshkova, N.V., Vaganov, E.A., Krutovsky, K.V.: Symmetry of siberian larch transcriptome. Journal of Siberian federal university **8(3)**, 278–286 (2015)
12. Sadovsky, M.G., Bondar, E.I., Putintseva, Y.A., Oreshkova, N.V., Vaganov, E.A., Krutovsky, K.V.: Seven-cluster structure of larch chloroplast genome. Journal of Siberian federal university **8(3)**, 268–277 (2015)
13. Sadovsky, M.G., Senashova, M.Y., Putintseva, Y.A.: Chloroplasts and Cytoplasm: Structure and Functions, chap. Chapter 2, pp. 25–95. Nova Science Publishers, Inc. (2018)
14. Sanguinetti, G., et al.: Gene regulatory network inference: an introductory survey. In: Gene Regulatory Networks, pp. 1–23. Springer (2019)
15. Satange, R., Chang, C.k., Hou, M.H.: A survey of recent unusual high-resolution DNA structures provoked by mismatches, repeats and ligand binding. Nucleic acids research **46(13)**, 6416–6434 (2018)
16. Wolstenholme, D.R.: Animal mitochondrial DNA: structure and evolution. International review of cytology **141**, 173–216 (1992)