Function vs. Taxonomy: The Case of Fungi Mitochondria ATP Synthase Genes

Michael Sadovsky[1,2(✉)], Victory Fedotovskaya[2], Anna Kolesnikova[2,3], Tatiana Shpagina[2], and Yulia Putintseva[2]

[1] Institute of Computational Modelling of SB RAS, Akademgorodok, 660036 Krasnoyarsk, Russia
msad@icm.krasn.ru

[2] Institute of Fundamental Biology and Biotechnology, Siberian Federal University, Svobodny prosp., 79, 660049 Krasnoyarsk, Russia
viktoriia.fedotovskaia@gmail.com, kolesnikova.denovo@gmail.com, shpagusa@mail.ru, yaputintseva@mail.ru

[3] Laboratory of Genomics and Biotechnology, Federal Research Center RAS, Krasnoyarsk, Russia
http://icm.krasn.ru

Abstract. We studied the relations between triplet composition of the family of mitochondrial *atp*6, *atp*8 and *atp*9 genes, their function, and taxonomy of the bearers. The points in 64-dimensional metric space corresponding to genes have been clustered. It was found the points are separated into three clusters corresponding to those genes. 223 mitochondrial genomes have been enrolled into the database.

Keywords: Order · Clustering · K-means · Elastic map · Stability · Evolution

1 Introduction

The problem of the interrelation of structure of nucleotide sequences, functions encoded in them, and taxonomy of their bearers still challenges researchers. A rapid growth of sequenced genetic data supports a progress in this problem. Yet, it is far from a completion, and the basic reason standing behind is the complexity of the phenomenon under consideration. Besides, one should keep in mind that the details of the problem statement may affect seriously both the answer, and the problem itself. In particular, one should define what is function, structure, and taxonomy, to get an exact, unambiguous and comprehensive answer on the question.

Here we try to reveal the interrelation and contribution of each entity, i.e. *structure*, *function* and *taxonomy* into their interplay and phenomenae observed in nature. Evidently, the answer depends strongly on the exact notion of what *structure* is, first of all. Luckily, the notion of a *function* is significantly less arguable, as well as the notion of *taxonomy*. The point is that the diversity and

I. Rojas et al. (Eds.): IWBBIO 2019, LNBI 11465, pp. 335–345, 2019.
https://doi.org/10.1007/978-3-030-17938-0_30

abundance of structure identified in nucleotide sequences is great enough (see, e.g. [1–6]), and those structure are quite different and may not be reduce one to another.

Everywhere below, **structure** is a frequency dictionary of triplets developed over some nucleotide sequence. We shall consider two types of triplet dictionaries, to be exact; they differ in the reading frame shift t. A triplet frequency dictionary W_3 is the set of all triplets $\omega_1 = \mathsf{AAA}$ to $\omega_{64} = \mathsf{TTT}$ together with their frequency

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

Here n_ω is the number of specific triplet ω observed over a sequence, and the reading frame (of the length 3) moves along a sequence with the step $t = 1$. Triplet frequency $\overline{W}_3$ is developed in the way similar to W_3, but for $t = 3$. The definition (1) must be changed then for

$$f_\omega = \frac{n_\omega}{M}, \tag{2}$$

where M is the total number of triplets counted within a sequence; obviously, M is three times less than N, for $\overline{W}_3$. Such frequency dictionaries have been used to reveal the relation between structure and taxonomy, see [7–9] for details. Further, we stipulate that there are no other symbols in genetic matter, but $\aleph = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$.

We studied 223 mitochondrial genomes of five fungal division: *Basidiomycota* (24 entries), *Ascomycota* (185 entries), *Blastocladiomycota* (2 entries), *Chytridiomycota* (6 entries) and *Zygomycota* (6 entries) were downloaded from NCBI GenBank. To reveal the interplay between all three issues mentioned above, we used the genes *atp*6, *atp*8 and *atp*9 belonging to ATP synthase genes family. The primary function of mitochondria is a production of energy via oxidative phosphorylation. In general, they encode 14 conserved protein-coding electron transport and respiratory chain complexes genes (*atp*6, *atp*8, *atp*9, *cob*, *cox*1, *cox*2, *cox*3, *nad*1, *nad*2, *nad*3, *nad*4, *nad*4L and *nad*6) and have no difference in function [10–12]. Using CLC Genomic Workbench v.10 we retrieved the annotations and the sequences of three standard mitochondrial protein encoding genes involved into the oxidative phosphorylation (these are *atp*6, *atp*8, *atp*9). Next, the sequence for each gene has been prepared in two versions:

(1) *gene* is a sequence containing exons and introns as it is presented in a genome, and
(2) *CDS (coding DNA sequence)* is a sequence free from introns, in fact, it corresponds to a mature RNA ready for protein translation.

Besides, ATP synthase genes are quite often used for phylogeny implementation [13–15].

As soon, as all the genes are isolated (in two versions each), the sequences have been transformed into the frequency dictionary W_3 or $\overline{W}_3$, respectively, with *ad hoc* software. Next, due to *VidaExpert*[1] freeware the distribution of

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

the points corresponding to genetic entities was analyzed. Transformation of sequences into frequency dictionaries allows to implement powerful and efficient tools of up-to-date statistical analysis and multidimensional data visualization.

1.1 Clustering Techniques

Clustering is the key tool of this research; we have used K-means and elastic map technique. K-means is well known and exhaustively described method of clustering, hence we shall not describe the method here in detail (see [16] for more details).

Also, the elastic map technique has been used to cluster and analyze data distribution, in triplets frequency space. Since this method is quite new, we describe it here in few details. To start, one must find out the first and the second principal components, and develop a plane over them (as on axes); next, each data point must be projected on this plane. Secondly, each data point must be connected to its projection with a mathematical spring. That latter has infinite expansibility and the elasticity coefficient remains permanent, for any expansion. Thirdly, figure out the minimal square comprising all the projections, and change it with the elastic membrane. That latter is supposed to be homogeneous, so that it may bend and expand. Next, release the system to reach the minimum of the total deformation energy. The elastic membrane would transform into a jammed surface, and this is the two-dimensional manifold approximating the data set. Fourthly, redefine each point on the jammed surface through the orthogonal projection. Finally, cut-off all the springs, so that the jammed surface comes back to a plane. That is the elastic map representing the cluster structuredness, if any, in the data set [17–19].

To identify clusters, we used the local density of points. That latter is defined as following. Supply each point of an elastic map (in so called inner coordinates, when the jammed surface is already flattened) with a bell-shaped function, e.g.

$$f(r) = \mathcal{A} \cdot \exp\left\{\frac{(r - r_j)^2}{\sigma^2}\right\}. \tag{3}$$

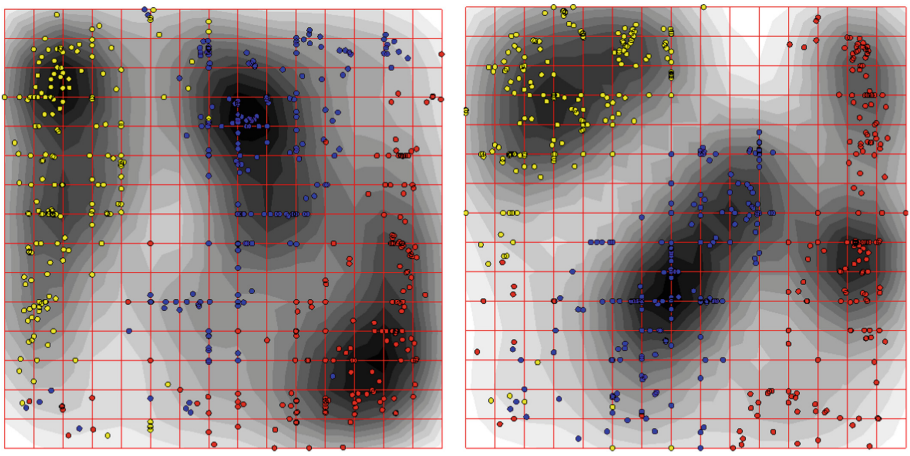
Here r_j is the coordinate vector of j-th point, and σ is an adjusting parameter (that is a specific width of the bell-shaped function). Then the sum function

$$F(r) = \mathcal{A} \cdot \sum_{j=1}^{N} \exp\left\{\frac{(r - r_j)^2}{\sigma^2}\right\}, \tag{4}$$

is calculated; the function $F(r)$ is then shown in elastic map.

2 Genes Distribution

We start from the clustering obtained due to elastic map technique and then consider the structuredness provided by K-means classification.



(a) (b)

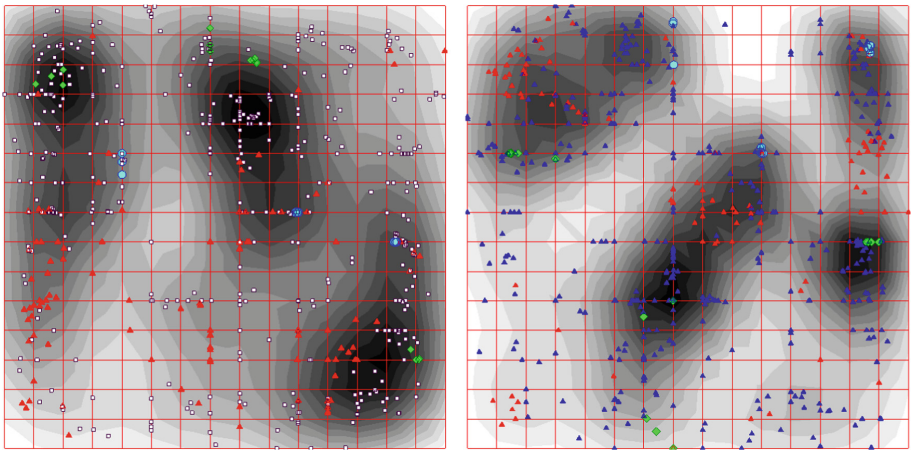
Fig. 1. Distribution of *atp* genes over elastic map. *atp*6 is in red, *atp*8 is in green and *atp*9 is in yellow; left is for W_3 and right is for $\overline{W}_3$ dictionaries. (Color figure online)

2.1 Elastic Map Clustering

Figure 1 shows the distribution of the genes on the soft elastic map (with 16×16 grid) and elasticity coefficients defined by default. Also, this figure shows the local density function, in grey scale. Figure 1(a) shows the distribution of genes in triplet frequency space obtained for W_3 dictionaries (developed for gene sequences with introns and exons). Evidently, there are three distinct clusters in this figure. Surprisingly, the clusters are gene specific, with a high accuracy: the left cluster gathers mainly *atp*9 genes (shown in yellow), the right cluster gathers mainly *atp*6 genes (shown in red) and the central one gathers mainly *atp*8 genes (shown in green). Or course, there are some "escapees": the genes that occupy an opportunistic cluster. The key point here is that this distribution is obtained for gene sequence (i.e. those with introns), and the reading frame shift $t = 1$.

Figure 1(b) shows similar distribution obtained for $\overline{W}_3$ dictionaries, with $t = 3$. In fact, these transformation into a triplet frequency dictionary completely corresponds to protein translation. There is no surprise in improved clustering observed for these dictionaries: the number of "escapees" goes down here.

The cluster structure shown in Fig. 1 is doubtless. Local density visualization technique makes it unambiguous. The clusters (in inner coordinates) are obviously isolated from each other. Coloring used to identify the peculiar gene type also unambiguously proves very high coherence of a cluster identified through triplet frequencies, and the gene type occupation. Of course, there are very few exceptions in the occupation: some genes join an opportunistic cluster. Nonetheless, the greatest majority of specific genes (say, *atp*6) tend to occupy the cluster that is identified through the triplet statistics, not with a functional role of a gene.



(a) (b)

Fig. 2. Distribution of species over elastic map. *Candida* spp. are in red, *Saccharomyces* spp. are in blue and *Fusarium* spp. are in green; left is for W_3 and right is for $\overline{W}_3$ dictionaries. (Color figure online)

Originally, the basic goal of our paper is to compare the impact of each entity from a triad *structure* – *function* – *taxonomy* on their common interplay pattern. To reveal the impact of each entity into this interplay, we checked the distribution of species in the patterns obtained through the clustering of frequency dictionaries W_3 and $\overline{W}_3$, respectively. Figure 2 shows such distribution, for three most abundant genera of fungi: *Candida* spp., *Fusarium* spp. and *Saccharomyces* spp. Again, Fig. 2(a) shows the distribution of W_3 triplet frequency dictionaries, and Fig. 2(b) shows the distribution of $\overline{W}_3$. Evidently, these three most abundant species are spread among the clusters rather equally; one may expect that other species are spread in similar manner, with obvious constraint coming from the finite (and small) number of some species comprising a genus.

2.2 *K*-means and Structure-Function Interplay

In Sect. 2.1, a direct evidence of the prevalence of function over the taxonomy is shown, for ATP synthase genes family of fungi mitochondria. Here we consider and analyze the structuredness obtained in the set of point corresponding to triplet frequency dictionaries due to K-means.

For each database (i.e. that one with W_3 dictionaries, and that one with $\overline{W}_3$ dictionaries) a classification through K-means has been developed; we implemented the classification for $K = 2$, $K = 3$, $K = 4$ and $K = 5$. Two issues must be kept in mind here: the former is stability of classification, and the latter is separability of classes. The first problem is immanent for K-means. Since a classification starts from a random allocation of the points into K classes, then there is no guarantee of the identity of the final configuration: it might change,

for different runs of the procedure. So, the idea of stability is to check whether a desirable number of runs converge to the same configuration, or not. If yes, then the classification is stable, otherwise it is unstable.

Unlike the elastic map technique, K-means does not yield the "natural" number of classes[2]; a researcher has to fix it at his own. On the other hand, one can develop a classification for various number of classes, and trace the transfer of elements of the classes, as they number grow up. This transfer is of special interest. We have developed the classifications for $2 \leq K \leq 5$ with K-means, for two issues: the former is taxonomy, and the latter is function.

Let us consider this point in detail. First, we consider the results of K-means implementation in terms of taxonomy. Figure 3 shows this series of four classifications as a layered graph; the classes are the nodes, and arrows are the edges. The edges indicate the transfer of the elements from a class to "younger" one (i.e., the transfer observed in two classifications with K-means and $K+1$-means). Complete layered graph is defined rather apparently: that is the layered graph where each node in K-th layer is connected to all nodes in $K+1$-th layer. In such capacity, the graph shown in Fig. 3 is almost complete: it has 25 edges, while the complete one must have 38 ones. In other words, the graph shown in this figure is far from a tree.

Figure 4 shows the graph observed for genes distribution. At the first glance, it looks pretty similar to that one shown in Fig. 3: it also has 22 edges (cf. to 28 in the graph shown in Fig. 3). Basic difference of that former consists in the abundances of objects (genes, in our case) and their preferences when transferred from node to node. This point is outlined with bold colored arrows connecting the specific nodes that comprise the genes with high predominance. The subgraph comprising the nodes and edges with high predominance of the genes makes a tree.

The composition of species in the graph shown in Fig. 3 is rather uniform: one can found any family in each node of the graph; in other words, the species tend to distribute themselves over the classes almost equally, so that no order or structuredness might be found. The pattern shown in Fig. 4 is drastically different; first of all, there are two isolated leaves in the graph. These are the classes with unique incident edge each, observed for $K = 5$; the former comprises *atp*8 genes (169 entries), and the latter comprises *atp*9 genes (86 entries). It should be stressed that *atp*8 genes differ, to some extent, from other ones, in this pattern: they are always comprised into a single cluster, for all $2 \leq K \leq 5$. The clusters enlisting *atp*8 genes also contain *atp*6 genes, for $2 \leq K \leq 4$. Evidently, the genes *atp*6 exhibits quite similar behaviour to *atp*8 genes: they tend to occupy the same cluster and split into two separated clusters only for $K = 5$, when *atp*8 comprise the isolated leaf in the graph, and *atp*6 comprise the cluster slightly deteriorated with other genes (14 entries of *atp*8 and 5 entries of *atp*9).

The family of *atp*9 genes is the only one tending to occupy a separate cluster, regardless the clustering technique used to identify the clusters. Such solidity

[2] An advanced version of K-means yields the maximal number of distinguishable classes, see Sect. 3.

Fig. 3. Transfer of the elements in a series of K-means classifications, $2 \leq K \leq 5$, for species distribution under classification with W_3 dictionaries.

remains even for a splitting of a cluster, as K exceeds 3. Indeed, a single cluster comprising[3] the genes of *atp*9 family splits into two clusters of approximately equal abundance, for $K = 4$ and $K = 5$. Still, there is no significant presence of the genes of other families in these two individual clusters.

[3] With respect to a minor deterioration by other genes.

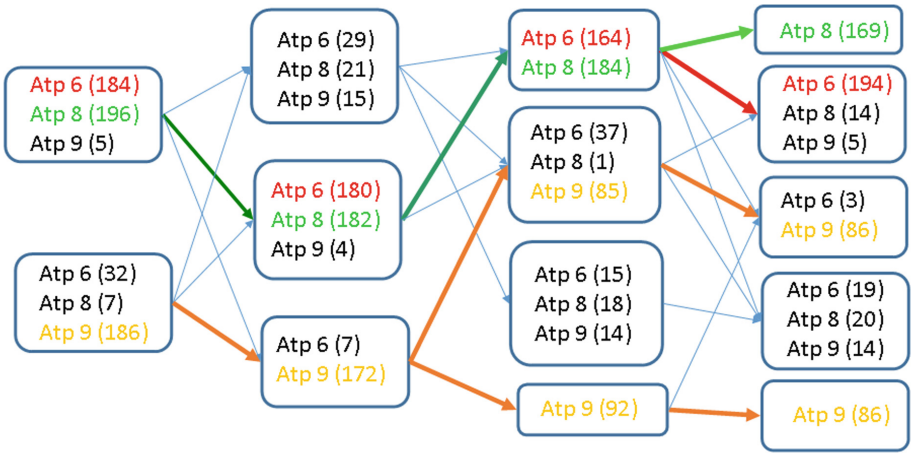


Fig. 4. Transfer of the elements in a series of K-means classifications, $2 \leq K \leq 5$, for genes distribution under classification with W_3 dictionaries developed over CDS. (Color figure online)

Let now consider in more detail the composition of two clusters comprising *atp*9 genes, as $K = 4$ (see Fig. 4). We examined the composition of these two clusters. It was found that all five divisions are splitted out between these two clusters almost homogeneously: the abundances of each specific division in a cluster is approximately proportional to the total abundance of this division in dataset. On the contrary, the genera are not split between these two clusters: it means that two genera belonging to the same division may occupy opportunistic clusters while the species belonging to a genus are mainly found in the same cluster, with a single exclusion. *Candida santjacobensis* is the only species found in the cluster comprising *atp*9 genes, only. All other genes of this genus (32 species) occupy the opportunistic cluster.

3 Discussion

Here we examined the mutual impact of three basic genetic entities (these are *structure*, *function* and *taxonomy*) on the pattern of their interplay. To do that, we created a database comprising the ATP synthase genes of fungal mitochondria, namely, *atp*6 genes. The genes were then converted into triplet frequency dictionaries, so that each gene is now represented as a point in 64-dimensional Euclidean space where the triplet frequencies are the coordinates. Then we checked whether an inner structuredness could be found in the set of such points, and the answer was positive. We have found that there exist three clusters identified with non-linear statistics (called elastic map technique); besides, other type of structuredness has been found through the implementation of linear classification technique (K-means).

At the next step, we examined all the clusters in terms of

(i) species composition, and
(ii) genes composition of each cluster.

The composition of the clusters has been checked regardless the identification technique used to figure them out.

Strong prevalence of gene (i.e. structure) in the cluster formation has been found, for all the clusters developed due to various clustering techniques; see Figs. 1 through 4. Such predominance is not self-evident, in advance. For example, paper [7] unambiguously proves the strong prevalence of taxonomy over the function, when studied over the entire mitochondrial genomes. One may expect that genes are stronger that taxonomy, while it is not evident in advance, for sure. The predominance of genes impact proves the superiority of function over taxonomy, in pattern formation within the triad *structure – function – taxonomy*. Nonetheless, this is not an ultimate proof; there are few questions to be answered to get a final evidence of the predominance mentioned above; let them list here:

(i) class (or clusters) distinguishability,
(ii) implementation of other metrics that Euclidean one,
(iii) stability of classification obtained with K-means, and
(iv) indexing of a database used to reveal the interplay, in terms of various taxa occurrence.

All these questions are rather technical than essential. Meanwhile, there are some more questions with hard biology standing behind.

3.1 CDS, $\overline{W}_3$ and Dimension Reduction

Previously, we presented structuredness observed in fungi mitochondrial ATP synthase genes through K-means implementation to classify triplet frequency dictionaries $\overline{W}_3$ developed over CDS of those genes. In fact, CDS is equivalent to mature RNA ready for translation; reciprocally, $\overline{W}_3$ frequency dictionary is the dictionary of the codons, not just common triplets, i.e. it contains the triplet occupying the positions corresponding to the reading frame at the translation process.

This fact allows to classify or cluster the genes in other space, with less dimension. Indeed, one can easily change the codon frequencies into the frequencies of corresponding amino acid residues. Since $\overline{W}_3$ comprises the codons, not the triplets, then the frequency of an amino acid residue is just the sum of the frequencies of all synonymous codons. This apparent and clear transformation results in the change of 64-dimensional Euclidean space for 21-dimensional one, where the frequencies of amino acid residues (plus *Stop* signal) are the coordinates.

3.2 Gene Family Selection

We have carried out the study of the mutual interplay of taxonomy, function and structure on the basis of ATP synthase genes of fungi mitochondria. Meanwhile,

it may take sense to extend the set of genes incorporated into a study: in particular, the oxidative phosphorylation involves the proteins encoded with some other genes, but *atp* family. Thus, an inclusion of the genes (*nad*1, *nad*2, *nad*3, *nad*4, *nad*4L, *nad*5, *nad*6, *cob*, *cox*1, *cox*2, *cox*3 and *cob*) both all together in isolated groups may bring a lot of new knowledge towards the relation within the triad *gene structure*, *function* and *taxonomy*.

3.3 Genome Selection

Similar reasoning as that one discussed in the above subsection addresses the choice of genomes to be considered for the analysis of the interplay in triad *structure – function – taxonomy*. There is no guarantee that the pattern with high prevalence of structure over taxonomy is observed always, regardless a genetic matter taken into consideration. Obviously, mitochondrion genomes are very good object for such kind of study: they are extremely homogeneous in the function encoded in it, the have a single chromosome, and are very well studied. Anyway, a universality of the observation done over these genomes must be verified through the examination of other genomes. Chloroplasts seem to be the second to none, in such capacity, for the same reasons: a single chromosome, perfect conservation of functions, good quality of sequencing and annotation.

Yet, a study of organella genomes may not be an ultimate proof of the pattern presented above. Some other genetic system must be involved into consideration, to approve it. All these issues fall beyond the scope of this paper and should be done in due time.

Acknowledgement. We are thankful to Reviewer whose remarks made the paper apparently better.

References

1. Molla, M., Delcher, A., Sunyaev, S., Cantor, C., Kasif, S.: Triplet repeat length bias and variation in the human transcriptome. Proc. Nat. Acad. Sci. **106**(40), 17095–17100 (2009)
2. Provata, A., Nicolis, C., Nicolis, G.: DNA viewed as an out-of-equilibrium structure. Phys. Rev. E **89**, 052105 (2014)
3. Qin, L., et al.: Survey and analysis of simple sequence repeats (SSRs) present in the genomes of plant viroids. FEBS Open Bio **4**(1), 185–189 (2014)
4. Moghaddasi, H., Khalifeh, K., Darooneh, A.H.: Distinguishing functional DNA words; a method for measuring clustering levels. Sci. Rep. **7**, 41543 (2017)
5. Bank, C., Hietpas, R.T., Jensen, J.D., Bolon, D.N.: A systematic survey of an intragenic epistatic landscape. Mol. Biol. Evol. **32**(1), 229–238 (2015)
6. Albrecht-Buehler, G.: Fractal genome sequences. Gene **498**(1), 20–27 (2012)
7. Sadovsky, M., Putintseva, Y., Chernyshova, A., Fedotova, V.: Genome structure of organelles strongly relates to taxonomy of bearers. In: Ortuño, F., Rojas, I. (eds.) IWBBIO 2015. LNCS, vol. 9043, pp. 481–490. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16483-0_47

8. 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.) IWBBIO 2016. LNCS, vol. 9656, pp. 455–464. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-31744-1_41
9. Sadovsky, M., Putintseva, Y., Zajtseva, N.: System biology of mitochondrion genomes. In: Qian, P.Y., Nghiem, S.V. (eds.) The Third International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies, pp. 61–66, Venice/Mestre (2011)
10. Basse, C.W.: Mitochondrial inheritance in fungi. Curr. Opin. Microbiol. **13**(6), 712–719 (2010)
11. Gray, M.W., Burger, G., Lang, B.F.: Mitochondrial evolution. Science **283**(5407), 1476–1481 (1999)
12. Bullerwell, C.E., Lang, B.F.: Fungal evolution: the case of the vanishing mitochondrion. Curr. Opin. Microbiol. **8**(4), 362–369 (2005)
13. Esser, C., et al.: A genome phylogeny for mitochondria among α-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. Mol. Biol. Evol. **21**(9), 1643–1660 (2004)
14. Davoodian, N., et al.: A global view of *Gyroporus*: molecular phylogenetics, diversity patterns, and new species. Mycologia **110**, 985–995 (2018)
15. Nadimi, M., Daubois, L., Hijri, M.: Mitochondrial comparative genomics and phylogenetic signal assessment of mtDNA among arbuscular mycorrhizal fungi. Mol. Phylogenet. Evol. **98**, 74–83 (2016)
16. Fukunaga, K.: Introduction to Statistical Pattern Recognition. Academic Press, London (1990)
17. Gorban, A.N., Zinovyev, A.Y.: Fast and user-friendly non-linear principal manifold learning by method of elastic maps. In: 2015 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2015, Campus des Cordeliers, Paris, France, 19–21 October 2015, pp. 1–9 (2015)
18. Gorban, A.N., Zinovyev, A.: Principal manifolds and graphs in practice: from molecular biology to dynamical systems. Int. J. Neural Syst. **20**(03), 219–232 (2010). PMID: 20556849
19. Gorban, A.N., Zinovyev, A.Y.: Principal manifolds for data visualisation and dimension reduction. In: Gorban, A.N., Kégl, B., Wünsch, D., Zinovyev, A.Y. (eds.) LNCSE, vol. 58, 2nd edn, pp. 153–176. Springer, Berlin, Heidelberg, New York (2007). https://doi.org/10.1007/978-3-540-73750-6