Non-Coding Regions of Chloroplast Genomes Exhibit a Structuredness of Five Types

Michael Sadovsky[1,2(✉)], Maria Senashova[1], Inna Gorban[2], and Vladimir Gustov[2]

[1] Institute of Computational Modelling of SB RAS, Akademgorodok 660036, Krasnoyarsk, Russia
{msad,msen}@icm.krasn.ru
[2] Institute of Fundamental Biology and Biotechnology, Siberian Federal University, Svobodny prosp., 79, 660049 Krasnoyarsk, Russia
inn.gorban@gmail.com, v.gustows@mail.ru
http://icm.krasn.ru

Abstract. We studied the statistical properties of non-coding regions of chloroplast genomes of 391 plants. To do that, each non-coding region has been tiled with a set of overlapping fragments of the same length, and those fragments were transformed into triplet frequency dictionaries. The dictionaries were clustered in 64-dimensional Euclidean space. Five types of the distributions were identified: ball, ball with tail, ball with two tails, lens with tail, and lens with two tails. Besides, the multi-genome distribution has been studied: there are ten species performing an isolated and distant cluster; surprisingly, there is no immediate and simple relation in taxonomy composition of these clusters.

Keywords: Order · Probability · Triplet · Symmetry · Projection · Clustering

1 Introduction

Non-coding regions in DNA sequences have been supposed to be a kind of an evolutionary junk; currently, it is a well knows fact that such regions play essential role in gene regulation, and in the genetic information processing, in general [1–6]. The role of non-coding regions is not absolutely clear yet, and a lot could be found behind them. The non-coding regions are found elsewhere, in a genome of any taxonomy level, including organelle genomes. Here we studied the non-coding regions of chloroplast genomes, following the way present in [7–12].

Previously, a seven-cluster pattern claiming to be a universal one in bacterial genomes has been reported and very elegant theory explaining the observed patterns was proposed [7,8,11]. Later, we have expanded the approach for chloroplast genomes [12,13]. Here se present some preliminary results of a study of

© Springer Nature Switzerland AG 2019
I. Rojas et al. (Eds.): IWBBIO 2019, LNBI 11465, pp. 346–355, 2019.
https://doi.org/10.1007/978-3-030-17938-0_31

statistical properties of non-coding regions of chloroplast genomes carried out under the methodology described above [11–13].

In papers [7–12] the difference in triplet composition determined for coding and non-coding regions has been established. Let now introduce more exact definitions and notions for further analysis. Consider a symbol sequence $\mathfrak{T}$ from four letter alphabet $\aleph = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$ corresponding to a (chloroplast) genome stipulating that $\mathfrak{T}$ has no other symbols but those indicated above. The sequences have been downloaded from NCBI bank (391 entities). Each sequence has been tiled with the set of intersecting fragments of the length L; the fragments located in a sequence with the step t. Next, for each genome every fragments were transformed into a triplet frequency dictionary W_3 (see Sect. 2). The transformation changed a fragment with a point in 64-dimensional Euclidean space, and the cluster structuredness of the points has been revealed and studied.

We aimed to check whether the fragments of each specific genome form a pattern where each separate genome is clustered more or less separately. Speaking in advance, the hypothesis both holds true, and it does not. More specifically, the triplet frequency dictionaries may not be separated by various clustering techniques; on the other hand, labeling each fragment with species reveals a non-random distribution of the points in Euclidean space. Moreover, an individual distribution of the fragments in the space reveals five types of the distribution. A study of a common distribution exhibits extremely unusual behaviour of ten genomes that form a kind of clearly and evidently separated dense cluster located very far from the main body of the points of other genomes.

2 Frequency Dictionaries

391 chloroplast genomes have been retrieved from NCBI bank. Each genome has been tiled with a set of (intersecting) fragments of the length $L = 603$ symbols; the fragments moved along a sequence with the step $t = 11$. It should be noticed that the length L is divisible by 3, but the step t is not; this choice of the parameters of tiling is not accidental. The idea standing behind this pattern of the tiling is described in detail in [8,11–13].

Next, each fragment was marked with the number of central nucleotide of that former. Following the annotation of a genome, we selected the fragments completely falling into non-coding regions. No overlaps to a coding region has been permitted. Then each fragment has been transformed into a triplet frequency dictionary. Formally, a triplet frequency dictionary W_3 could be defined ambiguously, in dependence on the reading frame shift. Indeed, let $\omega = \nu_1\nu_2\nu_3$ be a triplet, i.e. three symbols in $\mathfrak{T}$ standing next each other. Locate the frame identifying a triplet at the very beginning of $\mathfrak{T}$; move then the frame along $\mathfrak{T}$ with the step t and count all the triplets occurred within $\mathfrak{T}$. Counting the number of copies n_ω of each triplet ω, one gets the finite dictionary $\mathsf{W}_{(3,t)}$. Changing then the number of copies for their frequency

$$f_\omega = \frac{n_\omega}{N}\,, \qquad \text{where} \qquad N = \sum_{\omega=\mathsf{AAA}}^{\mathsf{TTT}} n_\omega\,, \tag{1}$$

one gets the frequency dictionary $W_{(3,t)}$. Obviously, one may use a frequency dictionary determined for an arbitrary t; we shall use the frequency dictionaries $W_{(3,3)}$ type.

2.1 Clustering

We used the freely distributed software *VidaExpert*[1] to analyze and visualize the distribution of the non-coding regions of genomes, both individually, and in a group. To do that, an ensemble of the fragments covering the non-coding regions corresponding to a genome has been arranged into a data base, and the distribution of the triplet frequency dictionaries has been studied, in the space of principal components of the ensemble. Also, a set of ensembles was arranged into a joint data base, with the same analysis technique applied for visualization.

3 Results

We examined 391 chloroplast genomes trying to identify a pattern of the tripe frequency dictionaries distribution, in the principal components space. Here present some preliminary results towards the patterns yielded by non-coding regions

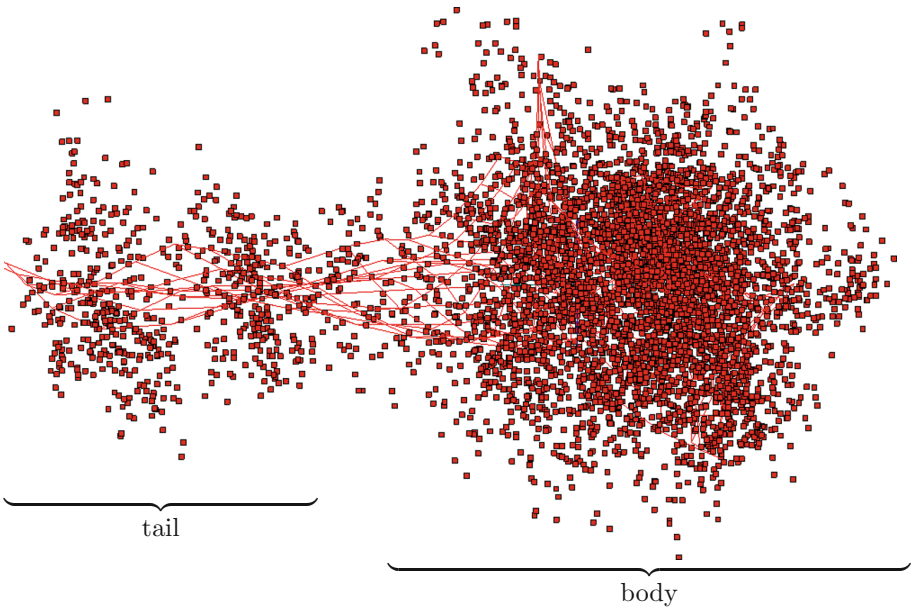
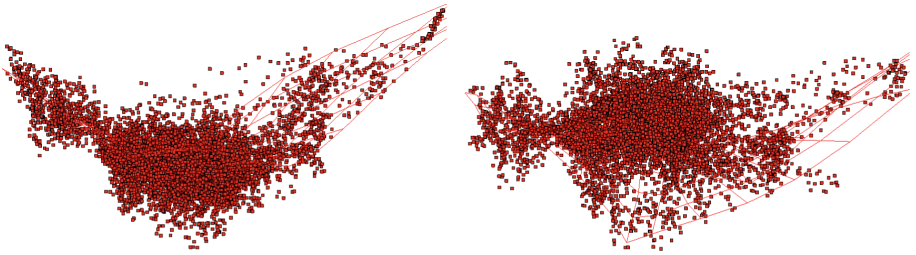


Fig. 1. Barley *Hordeum vulgare* subsp. *Spontaneum* chloroplast genome fragments distribution.

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

of chloroplast genomes, in the triplet frequency space. Subsection. 3.1 presents the results concerning the shape of the distribution observed over individual genomes, and Subsect. 3.2 presents similar results on the pattern observed for a mutual distribution of many genomes.

Let us also explain the terms *profile* and *above* used below to identify various projections. All figures provided below show the points distribution; that latter is a two-dimensional projection of a three-dimensional projection from 64-dimensional Euclidean space of triplet frequencies. All the figures present the distributions in three principal components (corresponding to the greatest, next and the third eigenvalue of the covariance matrix). *Profile* view means that the first principle components is located in the plane of a figure and directed from left to right; the second principal component here is also located in the plane, and directed from bottom to up. For *above* view the first principal components is located in the same way, but the second one orthogonal to the figure plane so that is looks out from the figure plane.



(a) (b)

Fig. 2. *Ricinus communis* chloroplast genome exhibits the lens with a tail structure. Left is *profile* view, and right is *above* view.

3.1 Individual Genome Clustering

Figure 1 shows a pattern to explain some terms used below. We classify the patterns in terms of *body* and *tail*: the patterns differ in the number of tails, and in the shape of a body. An examination of 391 genomes yielded five classes. These classes are:

(1) *Ball.* This is the pattern exhibiting no peculiar structuredness, the genome of *Erodium chrysanthum* is the typical representor (see Fig. 3(a)). This pattern differs from other ones due to a similitude of the distribution seen in various projections: any projection yields a ball. There are 7 genomes exhibiting this pattern.

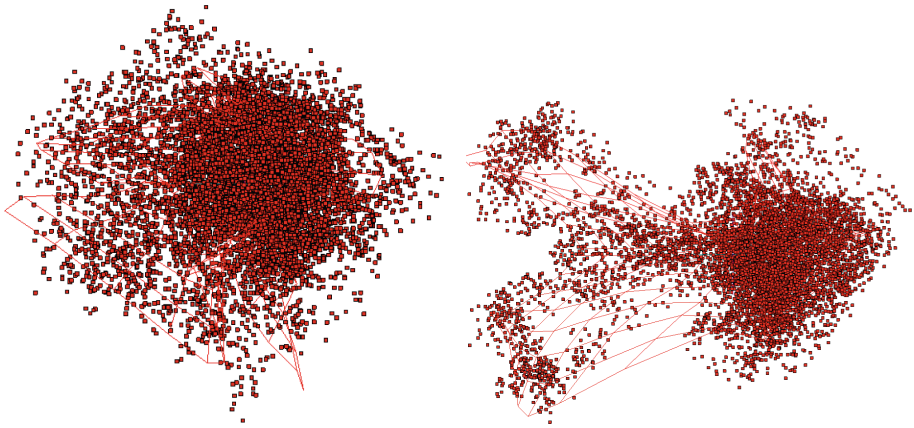
(2) *Ball and tail.* This is the pattern where the main body (*ball* is supplied with a clearly detectable other cluster (called *tail*); see Fig. 1 for details. The most surprising thing is that this tail looks like a (quite thick) ring, or torus; this is very unusual pattern, so the feasibility of minimum approximating

manifold must be provided properly [14]. There are 209 genomes exhibiting this pattern.

(3) *Ball and two tails.* This is the pattern resembling the previous one, while tail comprises two rings, not a single one. In such capacity, it might be called "scissors". There are 49 genomes exhibiting this pattern.

(4) *Lens and tail.* This pattern looks like a ball with tail (see Fig. 2(a)), in one projection, but in contrary to that former, it looks like a lens, or a ball segment, in other projection (see Fig. 2(b)). There are 45 genomes exhibiting this pattern.

(5) *Lens and two tails.* This pattern is similar to previous one, while it exhibits two tails, not a single one. There are 81 genomes exhibiting this pattern; see Fig. 4 for details.



(a) (b)

Fig. 3. *Erodium chrysanthum* chloroplast genome exhibits a ball-shaped structure (left), and *Liriodendron tulipifera* chloroplast genome exhibits a structure of ball with two tails (right).

Figures 1, 2 and 3 show all the structures observed in the family of 391 chloroplast genomes. The first question here arises whether those structures correlate to taxonomy of the genomes, or not. It should be noticed that the number of genomes exhibiting peculiar structure differs quite strongly, see the list of the structure above. In this Table, T means the total number of species in a division, L_1 (L_2, respectively) are the numbers of species within a division with *lens with tail* (*lens with two tails*, respectively) structures, B is the number

Table 1. Divisions distribution over the structure types; see text for details.

Division	T	L_1	L_2	B	B_2	B_1
Anthocerotophyta	1	0	0	0	1	0
Bryophyta	2	0	0	0	0	2
Marchantiophyta	3	0	0	0	3	0
Tracheophyta	385	45	81	7	45	207
Total	391	45	81	7	49	209

of species with *ball* structure, and B_2 (B_1, respectively) is the number of species within a division with *ball with tail* (*ball with two tails*, respectively) structure.

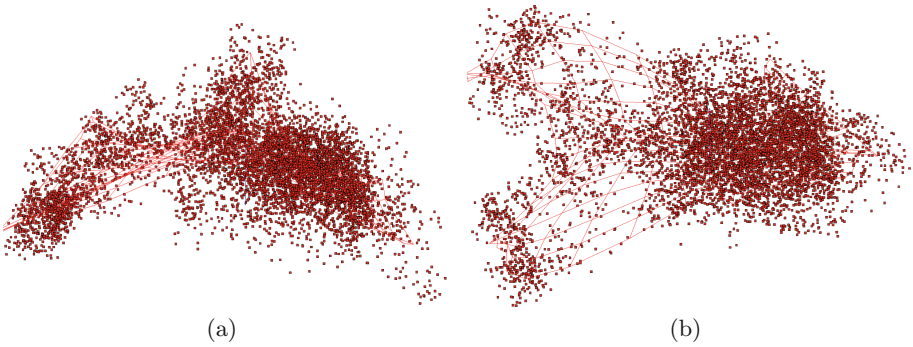


Fig. 4. *Lupinus luteus* chloroplast genome exhibits the lens with two tails structure. Left is *profile* view, and right is *above* view.

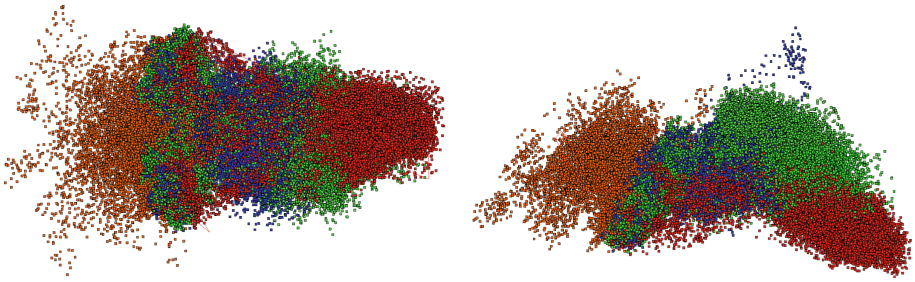
Table 1 shows the distribution of taxa at the division level. It should be said that the taxonomy composition of the divisions is quite biased: there are 6 or less species in three divisions; one hardly may expect to retrieve the taxonomy relation to a structure type over these division, due to a finite sampling effect. For *Tracheophyta* division is rather abundant and the distribution looks very far from a uniform one; besides, no other simple random distribution law might be fitted with these data (see Table 1).

3.2 Intergenomic Clustering

Previously, wonderful structuredness in bacterial genomes [7,8,11] has been reported. The structuredness manifests in clustering of considerable short fragments of a genome converted into triplet frequency arranged in seven clustering pattern, where six clusters represent coding regions of a genome, with respect to a reading frame shift, and the seventh one gathers fragments from non-coding regions. Later, this approach has been applied to a study of chloroplast genomes [12,13] and similar multi-cluster pattern has been found. The difference between bacteria and chloroplasts consists in different number of clusters observed in a pattern: bacteria genomes yield seven clusters, as maximum, while chloroplast ones yield up to eight clusters.

The structures mentioned above comprise the fragments identified both for coding and non-coding regions. In such capacity, the question arises whether one can reveal a relation between triplet composition, and taxonomy (for instance) of the genome bearers, in case of the comparison of a sufficiently abundant ensemble of genomes. Both for chloroplasts [10], and bacteria [9] the answer is positive: taxonomy may be traced in the system of clusters developed through K-means or other clustering techniques. The success of those researches has been

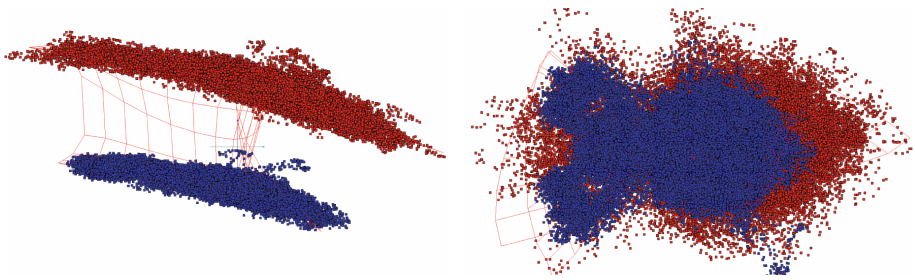
provided mainly by implementation of the entire genome into consideration, namely, coding and non-coding regions. So, the question arises whether similar relation between structure (namely, triplet composition) and taxonomy of the bearers, if non-coding regions are taken into consideration, only.



(a) (b)

Fig. 5. Simultaneous distribution of triplet dictionaries of non-coding regions of chloroplast genomes of five species.

Here we answer this question: yes, there is relation between taxonomy and triplet composition of the genome part comprising non-coding regions, solely. Figure 5 shows the simultaneous distribution of the fragments of non-coding regions converted into triplet frequency dictionaries of several species; to do it, we merged several data bases developed for individual genomes, into a single one and analyzed it. Different colors label different species; the cloud of the point belonging to the same species tend to form quite dense cluster, while these latter may not be separated with any unsupervised clustering technique. Figure 5(a) shows the view from above, and Fig. 5(b) shows the profile view of the distribution.



(a) (b)

Fig. 6. Ten genomes forming a distinct and outlying clusters; Fig. 6(a) shows the profile view, and Fig. 6(b) shows from bottom view.

3.3 Nine Mysterious Genomes

Nine genomes exhibit mysterious clustering behaviour: these are *Psilotum nudum*, AC AP004638, *Oryza sativa* Indica Group, AC AY522329, *Oryza sativa* Japonica Group, AC AY522330, *Panax ginseng*, AC AY582139, *Huperzia lucidula*, AC AY660566, *Helianthus annuus*, AC DQ383815, *Jasminum nudiflorum*, AC DQ673255, *Piper cenocladum*, AC DQ887677, *Pelargonium* × *hortorum*, AC DQ897681. These genomes form the distinct, apparent and clearly identified cluster that is located unexpectedly far from the main body formed by the other genomes. Figure 6 shows this clustering pattern. We have examined the behaviour of all these ten genomes, both separately and individually. It means that we checked the clustering structure formed by those genomes when combined with various number of other "normal" genomes."Normal" genomes form separately the cluster looking rather uniformly, from outer point of view. Those ten "escapees" also form the cluster that looks very uniformly from outer point of view. Meanwhile, together they exhibit the pattern where two clusters are evidently split and isolated one from other.

It should be said that the set of "normal" genomes is quite abundant: it comprises 381 genomes. Thus, we checked the separate cluster occurrence, for various less abundant subsets of "normal" genomes comprising up to 20 genomes and "escapees". It has been found that the "escapees" form the separated cluster in any combination of these latter, when compared to "normal" genomes.

4 Discussion

In papers [7,8,11] an approach to reveal a structuredness in bacterial genomes based on the comparison of frequency dictionaries $W_{(3,3)}$ of the fragments of a genome is presented; our results show that chloroplasts behave in other way. The always cluster in two coinciding triangles. The vertices of that latter correspond to phases of a reading frame shift and comprise the fragments with identical reading frame shift figure (reminder value). Moreover, unlike in [7,8,11], the chloroplast genomes exhibit a mirror symmetry.

Another important issue is that GC-content does not determine the positioning of the clusters, unlike for bacterial genomes. The pattern observed for bacterial genomes (triangle vs. hexagon) with central body comprising the non-coding regions of a genome is determined by GC-content. Both for bacteria [7,8,11] and chloroplasts [12,13], the fragments corresponding to non-coding regions of a genome always occupy the central part of a pattern; thus, the question arises towards a fine structure of those non-coding regions expressed in terms of statistical properties (and clustering) of the fragments falling purely into the non-coding regions. Here we present some preliminary results answering this question.

We analyzed non-coding regions separately from coding ones. First of all, the structuredness observed in non-coding regions differs significantly from that one observed over the whole genome. The patterns yielded by non-coding regions are more diffusive, in comparison to those observed for whole genome. Probably, the key difference consists in the lack of discernibility of the fragments belonging

to different species with unsupervised statistically based clustering technique. An inverse holds true: tracing the fragments belonging to the same species, one may see they comprise a dense and apparent cluster, if the distribution of the fragments belonging to different species is developed simultaneously.

Nonetheless, for each individual species the distribution of the fragments yields a specific pattern. We have identified five types of the distribution: *ball*, *ball with tail*, *ball with two tails*, *lens with tail* and *lens with two tails*. The structure called *ball with two tails* is the most surprising one: it ay not be approximated with good accuracy with a two-dimensional manifold of genus 0 (say, with a part of a plane, or hemisphere). On the contrary, the best starting manifold to approximate the pattern is a two-dimensional manifold of genus 2, i.e. a square with two holes in it.

Thus, we have proven an existence of a structuredness in the non-coding regions of chloroplast genomes; moreover, some relation to taxonomy of the bearers of the genomes may be traced. All the results show one can find a lot standing behind the simple statistical properties of non-coding regions of a genome, while more detailed study falls beyond the scope of this paper.

Acknowledgement. This study was supported by a research grant # 14.Y26.31.0004 from the Government of the Russian Federation.

References

1. Andolfatto, P.: Adaptive evolution of non-coding DNA in Drosophila. Nature **437**(7062), 1149 (2005)
2. Shabalina, S.A., Spiridonov, N.A.: The mammalian transcriptome and the function of non-coding DNA sequences. Genome Biol. **5**(4), 105 (2004)
3. Mercer, T.R., Dinger, M.E., Mattick, J.S.: Long non-coding RNAs: insights into functions. Nat. Rev. Genet. **10**(3), 155 (2009)
4. Kelchner, S.A.: The evolution of non-coding chloroplast DNA and its application in plant systematics. Ann. Mo. Bot. Gard. **87**(4), 482–498 (2000)
5. Guttman, M., Rinn, J.L.: Modular regulatory principles of large non-coding RNAs. Nature **482**(7385), 339 (2012)
6. Mattick, J.S., Makunin, I.V.: Non-coding RNA. Hum. Mol. Genet. **15**(suppl_1), R17–R29 (2006)
7. Gorban, A.N., Zinovyev, A.Y.: The mystery of two straight lines in bacterial genome statistics. Bull. Math. Biol. **69**(7), 2429–2442 (2007)
8. Gorban, A., Popova, T., Zinovyev, A.: Codon usage trajectories and 7-cluster structure of 143 complete bacterial genomic sequences. Phys. A: Stat. Mech. Appl. **353**, 365–387 (2005)
9. 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 Syst. Inf. Dyn. **7**(1), 1–17 (2000)
10. Sadovsky, M., Putintseva, Y., Chernyshova, A., Fedotova, V.: Genome structure of organelles strongly relates to taxonomy of bearers. In: Ortuño, F., Rojas, I. (eds.) Bioinformatics and Biomedical Engineering. Lecture Notes in Computer Science, pp. 481–490. Springer International Publishing, Cham (2015). https://doi.org/10.1007/978-3-319-16483-0_47

11. Gorban, A.N., Popova, T.G., Zinovyev, A.Y.: Seven clusters in genomic triplet distributions. Silico Biol. **3**(4), 471–482 (2003)
12. Sadovsky, M., Senashova, M., Malyshev, A.: Chloroplast genomes exhibit eight-cluster structuredness and mirror symmetry. In: Rojas, I., Ortuño, F. (eds.) Bioinformatics and Biomedical Engineering. Lecture Notes in Computer Science, pp. 186–196. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-319-78723-7_16
13. Sadovsky, M.G., Senashova, M.Y., Putintseva, Y.A.: Chapter 2. In: Chloroplasts and Cytoplasm: Structure and Functions, pp. 25–95. Nova Science Publishers, Inc. (2018)
14. Sadovsky, M.G., Ostylovsky, A.N.: How to detect topology of a manifold to approximate multidimensional data. In: Applied Methods of Statistical Analysis. Nonparametric Methods in Cybernetics and System Analysis, pp. 204–210, Novosibirsk, NSTU, NSTU PLC (2017)