Variation in the Structure of Cyanobacteria Genomes with the Length of the Sliding Window[*]

Vladimir Gustov[1][1111–2222–3333–599X], Maria Senashova[2][0000–0002–1023–7103]
and Mikhail Sadovsky[2][0000–0002–1807–0715]

[1] Siberian Federal University, 79 Svobodny st., Krasnoyarsk 660041, Russia
[2] Institute of Computational Modelling of the Siberian Branch
of the Russian Academy of Sciences, 50/44 Akademgorodok, Krasnoyarsk, 660036, Russia
{msen,msad}@icm.krasn.ru

Abstract. A seven-cluster pattern for bacterial genomes has been reported. The pattern is revealed through the distribution of formally identified short fragments of a genome converted into triplet frequency dictionaries. The pattern is found to be dependent on the parameters of the genome fragmentation, namely the length of a formally identified fragment within the genome, and the move step to identify the fragments to obtain an ensemble to reveal the clustering pattern. Here, we present the results of the studying the impact of the fragment length on the type and shape of the pattern.

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

1 Introduction

The diversity of structures found in biological macromolecules grows permanently. Their analysis is a key issue for researchers studying both classical biology and mathematics or computer science. Previously, some new structures were reported for bacteria [3, 2], transcriptomes [6, 8, 9] and chloroplasts [10, 5, 7]. The structure manifests in the mutual distribution of the formally identified fragments of a genome; to reveal the structure, one must convert the fragments into the triplet frequency dictionary $W_{(3,3)}$ and trace their distribution in the 64-dimensional metric space. Various fragments tend to gather into clusters, thus representing the structuredness.

The choice of the biological matter for such studies makes matter. Keeping aside of the sequencing, assembling, annotation etc., one faces an extremely high complexity of the objects under consideration. Prokaryotic organisms seem to be more convenient for such studies than eukaryotic ones. Prokaryotic genomes are significantly shorter and (almost always) consist of a single circular chromosome. Organelle genomes are even more advantageous when compared to prokaryotic ones in such capacity, since they encode the same function. Thus, there is no functional impact on the study of the interplay mentioned above, if a researcher studies the organelle genomes of the same group.

The papers mentioned above present the study of the inner structuredness for various genetic entities. The point is that the studies were carried out with the same parameters of tiling: the latter is the specific coverage of a genome with a set of (may be, overlapping) windows to identify the fragments for further clustering and revealing the inner structuredness.

Here we present some preliminary studies of the impact of variation in the tiling parameters on the structure revealed in the genome. Essentially, we check what happens with the structures reported previously, if the length of a tile increases. This growth means that one takes into consideration longer and longer correlations in the features found in a genome. Here, both features and correlations should be understood in broader sense.

2 Material and Methods

Tiling is a set of (overlapping) subsequences covering the genome under consideration. To do it, consider a genetic sequence of the length L from the four-letter alphabet $\aleph = \{A, C, G, T\}$. No other symbols are stipulated to occur in a sequence. To reveal the structuredness, tiling is developed. A window of the length Δ identifying a fragment within a sequence moves with the step t; we take $\Delta = 603$ and $t = 202$. Moving the window of the length Δ along the sequence to right with the given step t, one obtains the tiling. The latter is unambiguously determined by two parameters: Δ and t.

As soon as the tiling is developed, each tile is converted into the triplet frequency dictionary $W_{(3,3)}(j)$; here j denotes the tiles location along the sequence. Frequency dictionary is the list of all the triplets $\omega = \nu_1\nu_2\nu_3$ ranging from $\omega = AAA$ to $\omega = TTT$ provided with their frequency value f_ω. First, a sliding reading frame moving along the tile, captures the triplets $\omega = \nu_1\nu_2\nu_3$ counting the number of copies n_ω of each triplet. Next, the frequency of the triplet f_ω is derived by dividing the number n_ω by the total number of all the triplets:

$$f_\omega = \frac{n_\omega}{M}, \qquad (1)$$

where M stands for the total number of the triplets. They are counted along the fragment, with the reading frame shift equal to 3, thus providing neither gaps, nor overlaps. Further, we omit the subscript in the dictionary notation, unless it makes a confusion. The conversion of the sequence into an ensemble of the frequency dictionaries transforms it into a set of points in the 64-dimensional metric space.

For the purpose of the study, each point was labeled with the location coordinate j which is the number of central nucleotides of the relevant tile, 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 rigid coding region. To identify it we followed the annotation of the genome.

Seven labels of the index labeling are introduced: F_0, F_1, F_2, B_0, B_1, B_2 and J. The tile is indexed as F_k, $0 \leq \mathrm{k} \leq 2$, if the central nucleotide falls inside the coding region and the distance from the starting nucleotide of the coding region to the central one has the remainder k when divided by 3. This labeling holds for the genes located in the leading strand. Reciprocally, the tile is labeled with B_k, $0 \leq \mathrm{k} \leq 2$, 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 the opposite direction, since the genetic sequence is presented in the genetic bank with the leading strand only. Finally, the tile is indexed as J, if its central nucleotide falls out of the coding region.

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;
- the tiles indexed as F_2 and B_2 are colored in orange and yellow, correspondingly.

There is a point concerning the coloring scheme: it loses sense as the length of the fragment Δ grows. Indeed, the longer is the fragment, the greater is the number of the coding regions falling within the fragment. There is no reason to expect that all the newly coming coding regions observed within the increased fragment yield the same relative phase index. Thus, the fragment becomes a multi-phase entity, and the interfusion of the coding regions eliminates the unambiguity of the relative phase index. Nonetheless, we used the coloring system to relate the relative phase index to the very first coding region found in the fragment. Actually, it leads to mixing of the differently colored fragments within a cluster.

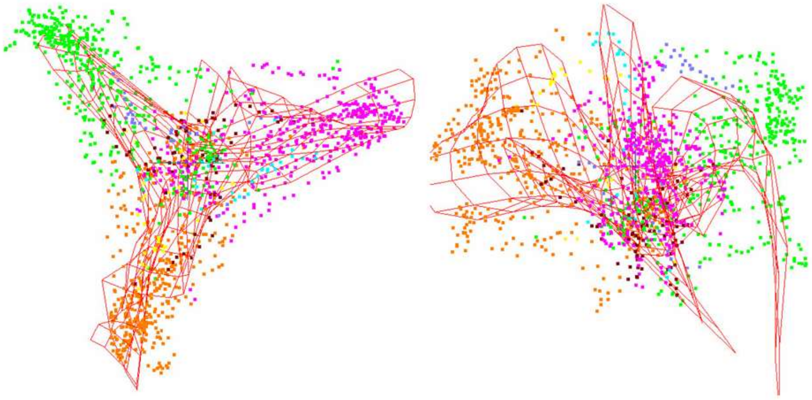

Fig. 1. Cyanobacteria genome structures observed for $\Delta = 603$. The projections on PC_1, PC_2 are shown in subfigs. 1(a), 1(c), 1(e) and 1(g); the projections on PC_2, PC_3 are shown in subfigs. 1(b), 1(d), 1(f) and 1(h). (See the genomes in the text).

We used the freely distributed software *VidaExpert*[1] to visualize the distribution of the tiles converted into the triplet frequency dictionaries in the metric space. We studied the distribution of the points corresponding to the tiling in the principal component space. To do this, we used the Euclidean metrics.

Any 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:

$$\sum_{\omega=\mathrm{AAA}}^{\mathrm{TTT}} f_{\omega} = 1 \qquad (2)$$

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

We studied 38 genomes of cyanobacteria downloaded from the EMBL–bank[2]. The studied species belong to the divisions Synechococcales, Nostocales, Chroococcales, Pleurocapsales and Oscillatoriales. Cyanobacteria are supposed to have a common ancestor with the existent chloroplasts, thus we tried to figure out relations between the structures observed in the latter, and those reported in chloroplasts [5, 7, 10].

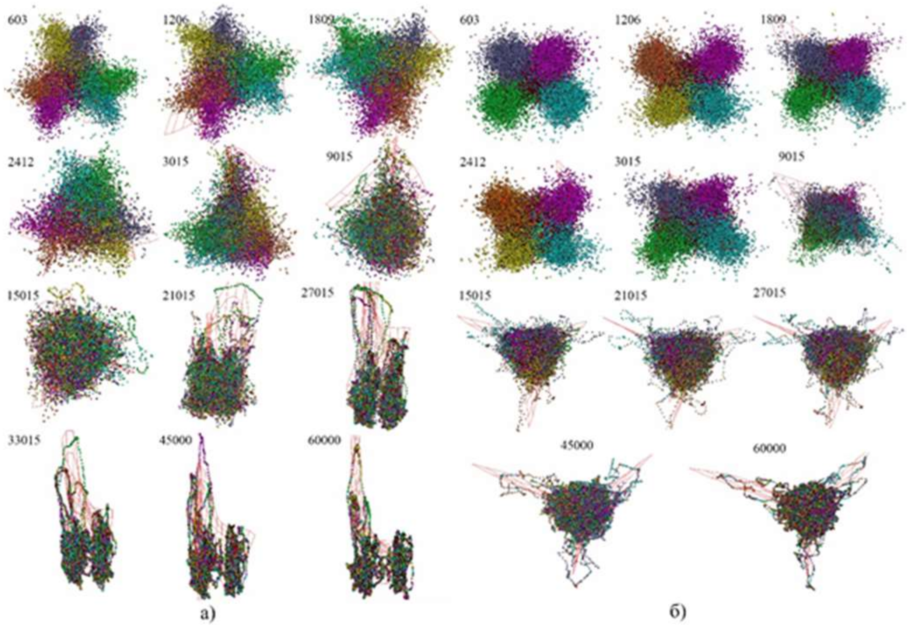

Fig. 2. Transformation of the structure, depending on Δ, for *Synechococcus sp.* WH 8102. Each subfigure presents two projections: the former is (PC_1, PC_2) (left), and the latter is (PC_2, PC_3) (right).

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

[2] https://www.ebi.ac.uk/genomes/bacteria.html

3 Results and Discussion

To begin with, let us concentrate on the structures observed for the "standard" parameters of tiling: that is $\Delta = 603$ and move step is $\mathrm{t} = 202$. Fig. 1 shows these patterns; subfigs. 1(a), 1(b) show Prochlorococcus sp. MIT 0604 (AC CP007753), subfigs. 1(c), 1(d) show Nostoc sp. PCC 7107 (AC CP003548), subfigs. 1(e), 1(f) show Synechococcus elongatus PCC 7942 (AC CP000100), and subfigs. 1(g), 1(h) show Synechococcus sp. WH 8109 (AC CP006882).

We analyzed the structures in the bacterial genomes mentioned above for Δ varying from 603 to 60 000 nucleotides; the move step t was equal to 202 or 201. The comparison of the structures observed with the extreme lengths of the fragment were of primary interest for us.

For $\Delta = 603$ and $\mathrm{t} = 202$ we observed four seven-cluster patterns, as described in [2, 3]; namely, these are:

- "parallel triangles" for AT-reach genomes (with GC-content close to 0.25), see Fig. 1;
- "orthogonal triangles" for the genomes with GC-content close to 0.35;
- "coinciding triangles" pattern characteristic of the genomes with GC-content close to 0.50, and finally
- a pattern degenerated into a plane with the six-beam cluster structure, for GC-content close to 0.60 (see Fig. 1).

Also, we investigated the variation of the pattern for the fragment lengths equal to 1206, 1809, 2412, 3015, 9015, 15015, 21015, 27015, 33015 and 45000 nucleotides. Fig. 2 illustrates the changes occurring in the patterns, as the fragment length grows. Evidently, the structure changes rather smoothly, as the window length increases. This change of the pattern seems to be universal for any initial structure.

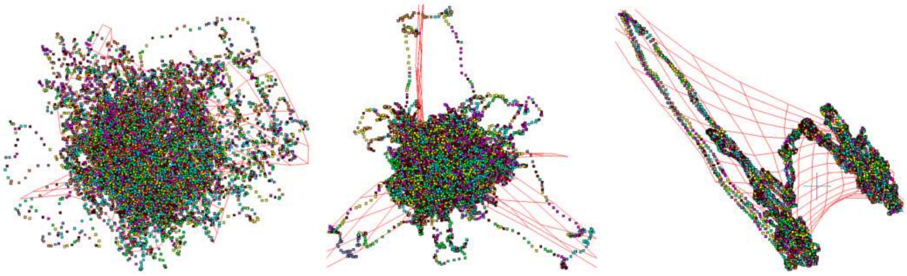

Fig. 3. Transformation Structures identified for $\Delta = 60\ 000$ for cyanobacteria; subfig. 3(a) shows *Synechococcus elongatus* PCC 6301; subfig. 3(b) shows *Anabaena variabilis* ATCC 29413 and subfig. 3(c) shows *Synechococcus sp.* CC9902.

We believe there exist 2 essentially different structures, and another substructure for $\Delta = 60\ 000$. These two essentially different structures are called ball (Fig. 3(a)) and Two clusters (Fig. 3(c)); the substructure mentioned above is called Snitch, see Fig.

3(b). The Ball structure resembles a ball, regardless of the initial type of the structure; this structure is characteristic of quite a chaotic distribution of the points corresponding to all six coding phases, and junk, over the ball. Careful examination shows that the structure looks like a set of threads; they can be well seen in the periphery of the ball. The Snitch substructure resembles, more or less, the ball structure. Three clearly identified threads resembling a protuberance characterize this substructure; they are extremely extended in size.

Table 1. Correspondence between the structure types for $\Delta = 603$ and $\Delta = 60\,000$; 2C stands for two clusters, S stands for the snitch and B stands for the ball.

	2C	S	B
Coinciding triangles	8	0	1
Six beams	2	0	3
Parallel triangles	0	0	2
Orthogonal triangles	0	5	17

The *Two clusters* structure consists of long threads which comprise not a ball, but two extended clusters connected with a bridge. The clusters differ in length, in all the cases. The clusters comprise the points belonging to two different non-overlapping parts of the genome: one may identify the points from each cluster (coloring, e. g.) and trace their location over the genome. Remarkably, the points of the same color occupy the sites over the genome separately, never mixing with each other.

Also, all these structures exhibit three (relatively independent) threads, in the principal component space; the thread composition depends on the divisibility of the window step t by 3. If the step is divisible, then the ball comprises a single thread; otherwise, three threads are observed (see Fig. 4).

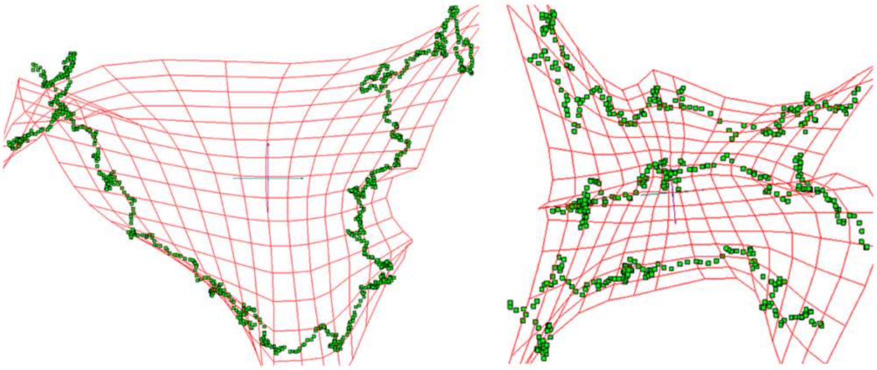

Fig. 4. *Acetobacter pasteurianus* genome structure observed for t divisible by 3 (Fig. 4(a)) and t indivisible by 3 (Fig. 4(b)). Here, the starting thousand of points are shown, to simplify visualization.

The points in the threads are located sequentially, if t is divisible by 3. Otherwise, the points in three threads observed for t indivisible by 3 are separated for each thread:

the threads comprise the fragments with the numbers giving the same remainder of the division of the location number of the fragment by 3, and the fragments within the thread are located sequentially, again.

Neither relation between the structure type of the genome of cyanobacteria observed for $\Delta = 60\,000$ and for $\Delta = 603$, nor taxonomy was found. Similarly, no relation of GC-content to these structures was revealed.

Table 2. Correspondence between the structure types for $\Delta = 603$ and $\Delta = 60\,000$.

	Ball	Snitch	Two clusters
Chroococcales	1	1	0
Pleurocapsales	1	0	0
Nostocales	4	3	0
Oscillatoriales	5	0	0
Synechococcales	13	1	10

Table 3 shows the GC-content values obtained at $\Delta = 60\,000$. It should be noted that the average GC-content for the two clusters structure exceeds that observed for other structures. However, it should be said that some genomes exhibiting these two structures other than two clusters have the GC-content values up to 60 %.

Table 3. Correspondence between the structure types of GC-content; σ is the standard deviation.

	$\mathrm{GC}_{\mathrm{min}}$, %	GC_{av}, %	$\mathrm{GC}_{\mathrm{max}}$, %	σ, %
Snitch	41.35	42.22	43.87	1.13
Ball	31.17	45.60	60.24	6.99
Two clusters	52.45	58.44	61.37	2.93

4 Conclusion

The majority of studies [12, 4, 11, 1] consider structures in terms of their functional role, or chemical properties. We focus exclusively on the statistical properties of biological macromolecules, regardless of their physical issues. The data presented here unambiguously prove that the inner structuredness in cyanobacteria genomes is observed at the window length $\Delta = 603$. Similar structuredness is observed for longer windows (up to $\Delta = 60\,000$), as well. The *Ball*, *snitch* and *two clusters* structures start to manifest themselves at $\Delta \approx 30\,000$ (see Fig. 2); further growth of the window length does not significantly change the structures.

Another important issue is the amazing impact of the triplet composition manifested at any length Δ of the window. Indeed, if the step t of the window move is divisible by 3, then a single thread pattern is observed; otherwise, one can see three threads.

References

1. Dittmann, E., Gugger, M., Sivonen, K., Fewer, D.P.: Natural product biosynthetic diversity and comparative genomics of the cyanobacteria. Trends in microbiology **23(10)**, 642–652 (2015)
2. 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
3. Gorban, A.N., Popova, T.G., Zinovyev, A.Y.: Four basic symmetry types in the universal 7-cluster structure of microbial genomic sequences. Silico Biology **5(3)**, 265–282 (2005), http://content.iospress.com/articles/in-silico-biology/isb00185
4. Kaneko, T., Tabata, S.: Complete genome structure of the unicellular cyanobacterium synechocystis sp. pcc6803. Plant and Cell Physiology **38(11)**, 1171–1176 (1997)
5. Sadovsky, M., Senashova, M., Malyshev, A.: Eight-cluster structure of chloroplast genomes differs from similar one observed for bacteria. ArXiv e-prints (Feb 2018)
6. 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)
7. 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. pp. 186–196. Springer International Publishing, Cham (2018)
8. 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(6)**, 278–286 (2015)
9. 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(6)**, 268–277 (2015)
10. 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)
11. Todorova, A.K., Juettner, F., Linden, A., Pluess, T., von Philipsborn, W.: Nostocyclamide: a new macrocyclic, thiazole-containing allelochemical from nostoc sp. 31 (cyanobacteria). The Journal of Organic Chemistry **60(24)**, 7891–7895 (1995)
12. Zervou, S.K., Kaloudis, T., Hiskia, A., Mazur-Marzec, H.: Fragmentation mass spectra dataset of linear cyanopeptides-microginins. Data in Brief p. 105825 (2020)