Unusual Distribution Structure of the Cyanobacteria Photosystem Genes in the Frequency Space of Triplets

Maria Yu. Senashova

Institute of Computational Modelling of the Siberian Branch of the Russian Academy of Sciences, 50/44 Akademgorodok, Krasnoyarsk, 660036, Russia

Abstract

Genes of the photosynthetic systems I and II are isolated for 45 cyanobacteria genomes. A frequency dictionary is built for each gene, to which a point in the 64-dimensional space of triplets is assigned. The photosystem gene structure in this space is considered. The genes are found to be clustered, depending on their belonging to the forward and reverse strands. Moreover, the points belonging to the forward and reverse strands form two perpendicular planes. The genes are grouped according to the type of bacteria within the main clusters. The values of the gene GC-content are distributed along the gradient.

Keywords

Order, distribution, clustering, evolution, triplets

1. Introduction

Photosynthesis, the conversion of solar energy into biomass, is one of the most fundamental processes on the Earth. Only photoautotrophic organisms such as cyanobacteria and plants can use photons to break down water molecules into hydrogen and molecular oxygen. The photosynthetic system of cyanobacteria, in contrast to purple and green bacteria, consists of two subsystems: photosystem I and photosystem II. The functions of these systems are mutually complementary. The primary function of photosystem II is to generate a strong oxidant, which initiates the oxidation of water and transfers its electrons to a membrane carrier. The primary function of photosystem I is to saturate these low-level electrons with energy in order to reduce NADP+. Since the energy of the total process is too great within the framework of one reaction center, there appeared two photosystems in the course of evolution, where different parts of this reaction occur. Their specific functions determine the features of their structure. Thus, photosystem I is symmetric, i.e. there two branches of electron transport, which makes it much faster. In contrast, photosystem II is asymmetric and has only one working branch, which slows down the transport of electrons, but makes it more controllable. Recently, significant progress has been made in determining the spatial structures of the photosystem of various cyanobacteria. However, they continue to attract the interest of researchers. The natural location of photosystem I (PSI), photosystem II (PSII), cytochrome (Cyt) b6f, and ATP synthase within thylakoid membranes at the molecular level was visualized. An inhomogeneous distribution of these four photosynthetic complexes was revealed and their dynamic features in a dense membrane environment were determined [1]. Structural studies of PS II of cyanobacteria led to the creation of a high-resolution spatial model of this huge complex. It was shown that the monomer of the photocomplex consists of 20 protein subunits, 54 pigment molecules, and 25 molecules of incorporated lipids. Mechanisms of operation of mobile electron carriers were proposed based on the structural data. The system for matching cluster atoms to the protein environment was described in detail [2]. New biochemical separation methods and structural characteristics of intermediate PSII complexes provided new insight into their protein composition and the spatial distribution of these

SibDATA 2021: The 2nd Siberian Scientific Workshop on Data Analysis Technologies with Applications 2021, June 25, 2021, Krasnoyarsk, Russia
EMAIL: msen@icm.krasn.ru (M. Senashova)
ORCID: 0000–0002–1023–7103 (M. Senashova)

complexes in the cell. The idea of the coordination of protein bonds and process of the PSII assembly was presented [3]. The role of the PsbU gene included in PSII was considered. It was found to be crucial for the stable architecture of the water-splitting system of the water separation system, which optimizes the efficiency of the oxygen production process [4]. For the structure of PSI and PSII of cyanobacteria, a comparison was made with the photosynthetic systems of higher plants, and assumptions on the evolution of the photosystem were made [5]. The functions and structure of the Psb27, Psb28, and Ycf48 hydrophilic assembly factors were discussed using structural, biochemical, and physiological information. A review was made of the role of these protein factors in the cyanobacterial assemblies of PSII, emphasizing their participation both in biogenesis and restoration of the photosystem from photodamage [7]. The new structures of PSI and PSII of cyanobacteria, algae, and plants shed light on the architecture and mechanism of the action of these complex membrane complexes and on the evolutionary forces shaping oxygen photosynthesis [8]. A review of the general structure of PSII was presented, followed by a detailed description of the specific structure of the catalytic center for water oxidation of the Mn_4CaO_5 cluster and its protein environment [10]. A comparison was made between phycobilisomes and PSII dimers. The most probable locations of terminal emitter subunits ApcD and ApcE inside lower cylinders of the PBS nucleus were determined, and chlorophyll PSII molecules collecting energy from PBS were identified [11]. In all these studies, the structure of the photosystem of cyanobacteria is considered from the viewpoint of biophysics and biochemistry. In the present work, the structure of the photosystem of cyanobacteria is considered from the viewpoint of bioinformatics, as the clustering of points is related to the genes of the photosystem in the frequency space of triplets.

2. Material and methods

In the present study, frequency dictionaries are used to represent genes as points in the frequency space of triplets. Let us describe in more detail the procedure for constructing dictionaries and identifying their structuredness. The genetic sequences of the length L, consisting of the alphabet symbols $\aleph = \{A, C, G, T\}$, are considered. In our case, the genes of the photosystem act as genetic sequences. For each of the sequences, a frequency dictionary of thickness 3 is compiled. The frequency dictionary of thickness 3 implies a list of all the triplets $\omega = \nu_1\nu_2\nu_3$ of consecutive nucleotides with an indication of the frequencies of these triplets. There can be 64 triplets in total. The frequency f_ω is the ratio of the number of copies n_ω of a given triplet to the total number of all the triplets N, where N is the sum of all n_ω:

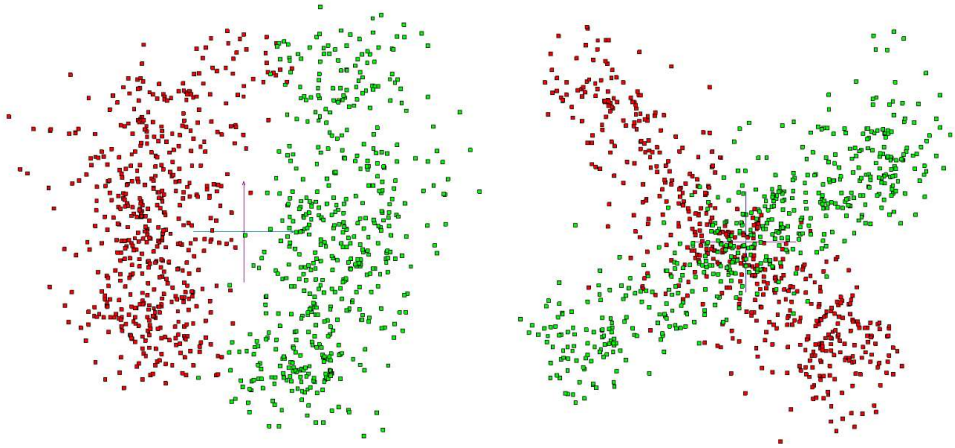
$$f_\omega = \frac{n_\omega}{N} \tag{1}$$

The triplets within a fragment are defined as follows: they do not intersect, but the combination of all the triplets completely covers the entire sequence in the gene. In other words, the reading frame in the construction of the dictionary is shifted by three nucleotides instead of one. In this case, the dictionary specifies the mapping of the genome into a 64-dimensional metric space. Two genes are considered to be close if the corresponding points in the 64-dimensional space are close in the sense of the Euclidean metric.

Thus, each gene is assigned a point in the 64-dimensional space of triplets. The following parameters are associated with each point: the name of the gene, the name of the species to which the gene belongs, the gene strand type (forward or reverse), the GC - content of the gene. The data view is built in the space of the first three principal components, calculated for the 64-dimensional space of triplets based on the obtained set of points in the VidaExpert program (http://bioinfo-out.curie.fr/projects/vidaexpert/). The projections of the space on the plane of the first and second as well as second and third principal components are considered.

3. Results and discussion

45 genomes of cyanobacteria from the EMBL bank have been examined, and the genes of the photosynthetic system have been annotated. The genes belonging to the forward and reverse strand have been found to form two clusters similar in the number of points belonging to them. That is, the number of the genes in the forward and reverse strand is approximately equal. Moreover, the points belonging to the same strand can be well approximated by an embedded plane. In addition, the planes corresponding to the points of each strand are perpendicular (Figure 1).



a) b)

Figure 1: The genes of the photosynthetic system I and II of cyanobacteria. The genes in the forward strand are shown in red, and those in the reverse strand are given in green. Fig. a) the view in the plane of the 1st and 2nd principal components; b) the view in the plane of the 2nd and 3rd principal components

At the same time, the genes of the same type do not form dense clusters, but they are extended along with the clusters of the forward and reverse strands (Figure 2).

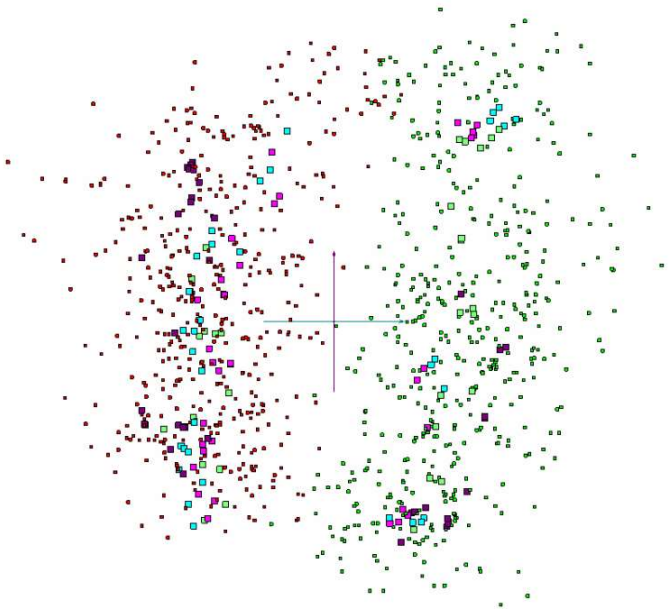


Figure 2: The location of the genes of the same type within the forward and reverse strand. For example, 4 genes are shown: crimson – *psaA*, turquoise – *psaB*, purple – *psbA*, light green – *psbB*. The view in the plane of the 1st and 2nd principal components

On the contrary, the genes belonging to cyanobacteria of the same species form clusters within the strands. Figure 3 shows the genes related to *Gloeobacter kilaueensis* (ID in the EMBL-bank CP003587) and *Gloeobacter violaceus* (BA000045), are denoted by crimson, *Synechocystis* (AP012205, AP012276, AP012277, AP012278, BA000022, CP003265) are marked in purple, *Nostoc* (CP003552, BA000019, CP003548, CP001037) are indicated in turquoise, *Prochlorococcus marinus* (CP000551, CP00552, CP000553) are indicated in light green.

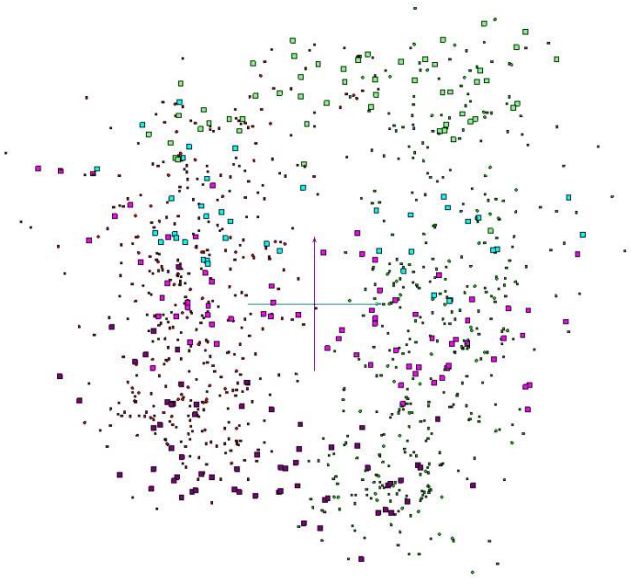


Figure 3: The location of the genes in the forward and reverse strand, colors are in accordance with the species of cyanobacteria. The view in the plane of the 1st and 2nd principal components

The distribution of the values of the gene *GC*-content has also been analyzed in the space of the first three principal components. It has been found that the values of the gene *GC*-content are located in the ascending order from the bottom (points indicated in green) upwards (points indicated in red) (Figure 4). The average value points are marked in yellow. There is a gradient distribution in the space of the frequencies of triplets for the values of the gene *GC*-content in the cyanobacteria photosystem. This type of distribution is common in complete genomes. In particular, it is found in the genomes of chloroplasts [9], in the genomes of mitochondria of higher plants, algae, mosses, lichens, fungi [6], and in *GC*-rich bacteria.

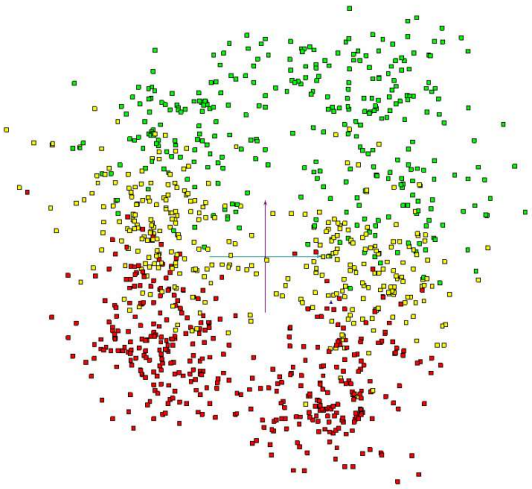


Figure 4: The distribution of the gene *GC*-content values in the plane of the first two principal components

4. Conclusion

The spatial structure of the genes in the photosynthetic systems of cyanobacteria in the space of triplet frequencies is similar to the structures previously studied for the complete genomes of chloroplasts, mitochondria, and bacteria in terms of the spatial distribution of the values of the gene GC -content. However, the difference from the spatial structure of the complete genomes is that a pronounced clustering of the genes of the forward and reverse strands is observed for the genes of the cyanobacteria photosystem. In addition, the points have been found to be grouped in the frequency space of triplets according to the types of organisms corresponding to them rather than according to the type of genes.

5. References

[1] S. Casella et al., Dissecting the native architecture and dynamics of cyanobacterial photosynthetic machinery, Molecular plant 10(11) (2017) 1434–1448. doi:10.1016/j.molp.2017.09.019.

[2] A.G. Gabdulkhakov, M. V. Dontsova, Structural studies on photosystem II of cyanobacteria, Biochemistry 78(13) (2013) 1524–1538. doi:10.1134/S0006297913130105.

[3] S. Heinz et al., Analysis of photosystem II biogenesis in cyanobacteria, Biochimica Et Biophysica Acta (BBA)–Bioenergetics .1857(3) (2016) 274–287. doi:10.1016/j.bbabio.2015.11.007.

[4] N. Inoue–Kashino et al., PsbU provides a stable architecture for the oxygen–evolving system in cyanobacterial photosystem II, Biochemistry. 44(36), 12214–12228 (2005). doi:10.1021/bi047539k.

[5] N. V. Karapetyan, Photosystem I of cyanobacteria: organization and functions, Advances in biological chemistry 41 (2001) 39–76.

[6] R. Kosarev, M. Senashova, M. Sadovsky, Intrinsic Structuredness of Mitochondria Genomes, in: L. Nozhenkova, T. Penkova, A. Korobko (Eds.), The 1st Siberian Scientific Workshop on Data Analysis Technologies with Applications 2020, volume 2727 of SibDATA'20, CEUR, 2020, Krasnoyarsk, Russia, pp. 66–74.

[7] P. D. Mabbitt, S. M. Wilbanks, J. J. Eaton–Rye, Structure and function of the hydrophilic Photosystem II assembly proteins: Psb27, Psb28 and Ycf48, Plant Physiology and Biochemistry 81 (2014) 96–107. doi:10.1016/j.plaphy.2014.02.013.

[8] N. Nelson, C. F. Yocum, Structure and function of photosystems I and II, Annu. Rev. Plant Biol. 57 (2006) 521–565. doi:10.1146/annurev.arplant.57.032905.105350.

[9] M. G. Sadovsky, M. Yu. Senashova, A. V. Malyshev, Amazing symmetrical clustering in chloroplast genomes, BMC Bioinformatics 21(Suppl 2) 83 (2020). doi:10.1186/s12859-020-3350-z.

[10] J. R. Shen, The structure of photosystem II and the mechanism of water oxidation in photosynthesis, Annual review of plant biology 66 (2015) 23–48. doi:10.1146/annurev-arplant-050312-120129.

[11] D. V. Zlenko et al., Coupled rows of PBS cores and PSII dimers in cyanobacteria: symmetry and structure, Photosynthesis research 133(1) (2017) 245–260. doi:10.1007/s11120-017-0362-2.