ISSN 2348-6201

Analysis of Dinucleotide Bias and Genomic Signatures Across Cyanobacterial Genomes

Ratna Prabha[1] and Dhananjaya P. Singh[1,2,*]
[1]Mewar University, Gangrar, Chittorgarh, Rajasthan-312 901
ratnasinghbiotech30@gmail.com
[2]National Bureau of Agriculturally Important Microorganisms, Kushmaur, Maunath Bhanjan 275103 INDIA
dpsfarm@rediffmail.com (*corresponding author)

Abstract

Dinucleotide relative abundance or frequency of dinucleotides in particular nucleotide sequence is reported as genomic signature as it is specific across different DNA samples and able to identify the variance among different groups whereas identical or similar for closely related organisms. Dinucleotide relative abundance value is identified to be overall constant due to numerous factors as dinucleotide stacking energy and DNA helicity, mechanisms of replication, repair and context-dependent mutation pressures and includes information about genome-wide processes. In this study, we analyzed genome sequence of 41 cyanobacteria to gain an insight of dinucleotide bias in their genomes. Across different dinucleotides, TA is broadly underrepresented followed by CG and then AC+GT whereas; AA+TT and GC occupied a major portion of dinucleotide distribution across all cyanobacterial genomes. Underrepresentation of TA seems to be influenced by GC-content of the members as it tends to decrease when there is an increase in GC content. Members with similar GC content possess similar pattern of genomic signature. Habitats also seem to influence the dinucleotide relative abundance values of the organisms because it is suggested that marine organisms.

Indexing terms/Keywords

Cyanobacteria; dinucleotide relative abundance; genomic signature; dinucleotide bias.

Academic Discipline And Sub-Disciplines

Biology.

SUBJECT CLASSIFICATION

Bioinformatics.

TYPE (METHOD/APPROACH)

Research Article.

Council for Innovative Research
Peer Review Research Publishing System
Journal: JOURNAL OF ADVANCES IN BIOTECHNOLOGY
Vol. 3, No. 3.
www.cirjbt.org , jbteditor@gmail.com

Due to these facts, dinucleotide relative abundance is depicted as **genomic signatures** that tend to be specific across different DNA samples and were found efficient to explain the variance across the DNA sequences of prokaryotes and eukaryotes including viruses and hosts and also provide information on their variation at codon sites [2-3]. Dinucleotide relative abundance value is found to be consistent throughout the genome and involves contribution of genome-wide processes such as replication, recombination and repair in it. Environmental factors such as ecology (e.g. energy sources and systems), temperature extremes, g-radiation damage, osmolarity gradients along with transfer of genomic DNA between organisms (either directly or indirectly) concurrently imposes impact on the genomic signature [1]. Dinucleotide relative abundance differentiate specifically imitate structural features of DNA such as duplex curvature, supercoiling etc. [4]. Dinucleotide relative abundance founds its root in dissimilarity measures calculated from dinucleotide counts and is also utilized for assessing evolutionary distances between homologous sequences as an alignment-free approach computation. Phylogenetic analysis on the basis of dinucleotide relative abundance distance (or "delta-distance") is specifically useful for whole genomes and provides logically sound results [5-7].

Dinucleotide relative abundance profile is found to be very stable and consistent throughout the genome even when only 50 kb fragments are considered [2, 4]. This stability is a resultant of many factors like limitation on dinucleotide stacking energy and DNA helicity, mechanisms of replication and repair and context-dependent mutation pressures [1, 4, 5, 8, 9]. The genome signature is also able to identify putative horizontally transferred DNA as it is typical for a given bacterial genome. Due to its species-specific character, this genomic signature allows recognition of anomalous genomic regions [10, 11, 12].

Special attention was provided towards studying prokaryotic genomes for analysis of biases in nucleotide composition and organization along with short oligonucleotide combinations held there in [13-15]. Much emphasis is given towards analyses of dinucleotide frequencies and codon usage [10, 16-19]. In archea and bacteria, usage of oligonucleotides are related to multiple properties as DNA base-stacking energy, codon usage and DNA structural conformation. Further, it is reported that prokaryotic DNA tends to be correlated in short range and information is encoded in short oligonucleotides. As compared to AT-rich and host-associated genomes, oligonucleotide usage vary more in GC-rich and free-living genomes [20].

Cyanobacteria (blue-green algae) represent one of the eleven major eubacterial phyla and extremely diverse group of prokaryotes in terms of their physiological, morphological and developmental characteristics. They are ancient group of photosynthetic prokaryote with a great distinction in term of their habitats, cellular differentiation strategies and physiological capacities [21]. In the last decades, increased technological developments in DNA-sequencing have facilitated sequencing of a number of cyanobacterial genomes comprising different physiological groups and species. Complete genome sequences of group of microbes including cyanobacteria allow a close inspection of genomic features and characteristics within and between different species.

This study was carried out to analyse dinucleotide frequencies and average absolute dinucleotide relative abundance difference across different cyanobacterial genomes.

Materials and Methods:

Dataset:

Entire dataset contains 41 different group of cyanobacteria for whom complete genome sequence is available. Complete genome sequences of 40 cyanobacteria were obtained from the NCBI ftp site (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). For *Arthrospira platensis* NIES-39, entire genome sequence was downloaded from GenBank database while entire gene sequences were downloaded from DOGAN database (http://www.bio.nite.go.jp/dogan/top). All 41 cyanobacteria belong to 5 different orders i.e. Chrococcales, Prochlorales, Nostacles, Oscillatoriales and Gloeobacterales. Chrococcales is largest

order having 22 members sequenced, whereas next one is Prochlorales with 12 members. Order Nostacles, Oscillatoriales and Gloeobacterales have 4, 2 and 1 member sequenced respectively (Table 1).

Calculation of dinucleotide relative abundance value

We determined the dinucleotide relative abundance value for each of the 41 cyanobacteria using the following equation:

$$\rho^*_{XY} = f_{XY} / (f_X f_Y)$$

where f_{XY} denotes the frequency of dinucleotide XY and f_X and f_Y denote the frequencies of X and Y, respectively. Program Count Motifs (http://kirillkryukov.com/study/tools/count-motifs/) was used to calculate the dinucleotide relative abundance values. We followed refined criteria of discrimination, proposed in earlier studies [9, 22] i.e. overrepresentation is indicated by + ($1.23 \leq \rho^*_{XY} < 1.30$), ++ ($1.30 \leq \rho^*_{XY} < 1.50$) and +++ ($\rho^*_{XY} \geq 1.50$), while underrepresentation is indicated by – ($0.70 < \rho^*_{XY} \leq 0.78$), – – ($0.50 < \rho^*_{XY} \leq 0.70$) and – – – ($\rho^*_{XY} \leq 0.50$).

Calculation of average absolute dinucleotide relative abundance difference

The dissimilarities in relative abundance of dinucleotides between two sequences (f and g) were calculated from Genome signature comparisons (δ*-differences) (webserver http://www.cmbl.uga.edu/software/delta-differences.html). This webserver computes δ*-differences using the following equation:

$$\delta^*(f,g) = 1/16 \sum |\rho^*_{XY}(f) - \rho^*_{XY}(g)| \text{ (Karlin et al, 1997)}$$

Program first divides each genome to non-overlapping segments of ~50,000 bp, then calculates the δ* value for each pair of segments from the two genomes, and gives the average of all comparisons between 50 kb segments multiplied by 1000 for convenience .

Statistical analysis

Statistical analysis i.e. calculation of mean, standard deviation and correlation analysis was carried out with SPSS 16.0 software.

Results

Comparison of dinucleotide relative abundance values across genomes

The distribution pattern of the frequencies of 16 dinucleotides i.e. symmetrized 10 dinucleotides of 41 species of cyanobacteria is shown in Table 1. Our study indicated that TA is broadly underrepresented followed by CG and then AC+GT as shown earlier [4, 23, 24]. Slight variation is observed in distribution pattern of CC+GG followed by TG+CA. Particularly for these two set of dinucleotides, it iswas observed that they followed an average to overrepresented distribution across all cyanobacteria (Table 1). CC+GG was found in higher occurrence in members of order Prochlorales and across Cyanothece species. AA+TT and GC occupied a major portion of dinucleotide distribution across all members of dataset. Underrepresentation of TA was observed to be influenced by GC-content of the members as it tended to decrease when there is an increase in GC content. Members with low GC content showed underrepresentation of CG (Table 1). Frequency range for TA was found to be highest followed by CG. Both of these dinucleotides show their distribution in a wide range which was also evident from Table 1. Wide distribution range was also observed for CC+GG, TG+CA and GC. Rest of the dinucleotides which generally involved combination of one strong nucleotide and one weak nucleotide (AC+GT, AG+CT, TC+GA) with the exception of AA+TT and AT showed narrow range of frequency as compared to rest of the dinucleotides (Figure 1).

ISSN 2348-6201

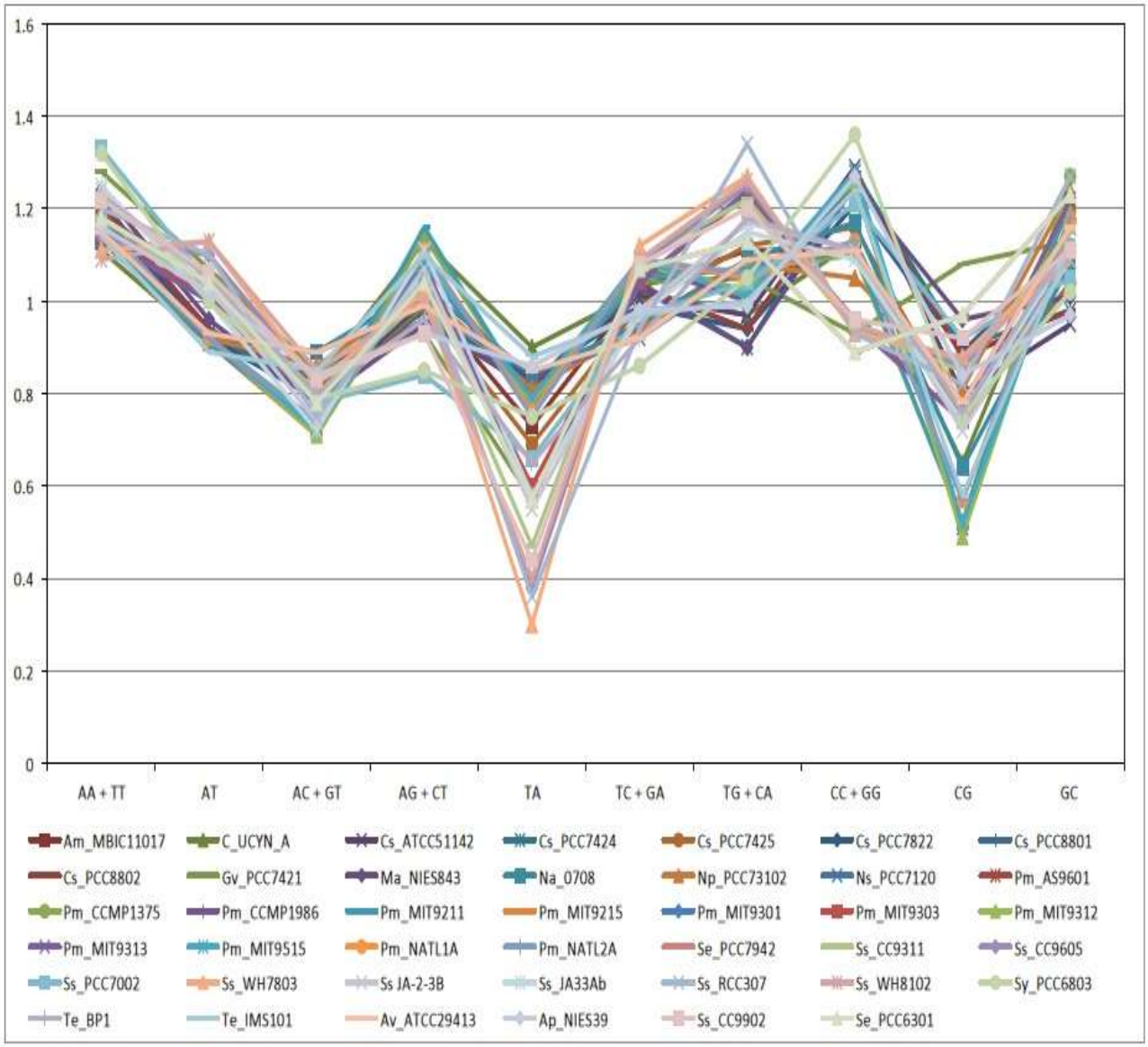

Figure 1. Distribution of dinucleotides across all the 41 cyanobacteria.

ISSN 2348-6201

Table 1. Distribution pattern of dinucleotide frequency in 41 cyanobacterial genomes. Colour coding varies in order red-yellow-green as 1.36 1 0.3. Overrepresentation is indicated by + ($1.23 \leq \rho^*_{XY} < 1.30$), ++ ($1.30 \leq \rho^*_{XY} < 1.50$) and +++ ($\rho^*_{XY} \geq 1.50$), while underrepresentation is indicated by – ($0.70 < \rho^*_{XY} \leq 0.78$), – – ($0.50 < \rho^*_{XY} \leq 0.70$) and – – – ($\rho^*_{XY} \leq 0.50$) (Karlin et al, 1997).

Taxonomy	Organism	Abbreviation used for Organism Name	GC %	AA + TT	AT	AC + GT	AG + CT	TA	TC + GA	TG + CA	CC + GG	CG	GC
Chroococcales	*Acaryochloris marina MBIC11017*	Am_MBIC11017	46.99	1.14	1.04	0.79	1.01	- 0.73	1.03	1.11	1.11	- 0.76	1.09
	Cyanothece sp. ATCC 51142	Cs_ATCC51142	37.97	1.19	0.92	0.84	0.98	0.82	1.01	0.97	+ 1.29	0.8	0.95
	Cyanothece sp. PCC 7424	Cs_PCC7424	38.5	1.21	0.93	0.79	0.98	0.82	1.04	0.9	+ 1.28	0.9	0.98
	Cyanothece sp. PCC 7425	Cs_PCC7425	50.66	1.19	1.03	0.79	0.99	-- 0.69	0.99	1.12	1.16	- 0.74	1.04
	Cyanothece sp. PCC 7822	Cs_PCC7822	39.87	1.22	0.92	0.79	1	0.83	0.99	0.94	1.21	0.89	1.12
	Cyanothece sp. PCC 8801	Cs_PCC8801	39.8	1.21	0.94	0.81	0.96	0.81	1.03	0.94	+ 1.27	0.88	0.98
	Cyanothece sp. PCC 8802	Cs_PCC8802	39.8	1.21	0.94	0.81	0.96	0.81	1.03	0.94	+ 1.27	0.88	0.98
	Synechococcus elongatus PCC 7942	Se_PCC7942	55.46	1.18	1.04	- 0.78	1.04	-- 0.57	1.07	1.13	0.89	0.97	+ 1.23
	Synechococcus sp. CC9311	Ss_CC9311	52.4	1.18	1.02	0.8	1.01	--- 0.47	1.09	1.22	0.94	0.85	1.15
	Synechococcus sp. CC9605	Ss_CC9605	59.2	1.14	1.1	0.82	1.01	--- 0.38	1.08	+ 1.25	0.95	0.86	1.11
	Synechococcus sp. JA-3-3Ab	Ss_JA33Ab	60.2	+ 1.25	1	- 0.72	1.11	-- 0.58	0.97	1.15	1.09	- 0.74	1.12
	Synechococcus sp. PCC 7002	Ss_PCC7002	49.16	++ 1.33	1.04	-	0.84	--	0.97	1.04	1.21	0.92	1.05

						0.78		0.66					
	Synechococcus sp. RCC307	Ss_RCC307	60.8	1.2	1.1	- 0.75	1.05	--- 0.36	0.94	++ 1.34	0.93	0.82	+ 1.27
	Synechococcus sp. WH 7803	Ss_WH7803	60.2	1.11	1.13	0.82	1.02	--- 0.3	1.12	+ 1.27	0.93	0.88	1.11
	Synechococcus sp. WH 8102	Ss_WH8102	59.4	1.09	1.13	0.85	1	--- 0.4	1.09	+ 1.26	0.95	0.87	1.09
	Synechocystis sp. PCC 6803	Sy_PCC6803	47.35	++ 1.32	1	0.79	0.85	- 0.75	0.86	1.05	++ 1.36	- 0.75	1.02
	Microcystis aeruginosa NIES-843	Ma_NIES843	42.3	+ 1.24	0.96	- 0.78	0.95	0.82	1.03	0.9	+ 1.25	0.96	1.01
	Thermosynechococcus elongatus BP-1	Te_BP1	53.9	+ 1.23	1.05	0.8	0.96	-- 0.65	0.92	1.18	1.11	- 0.76	1.13
	Synechococcus sp. CC9902	Ss_CC9902	54.2	1.22	1.06	0.83	0.93	--- 0.44	1.08	1.2	0.96	0.92	1.11
	Synechococcus elongatus PCC 6301	Se_PCC6301	55.5	1.18	1.04	- 0.78	1.04	-- 0.57	1.07	1.13	0.89	0.97	+ 1.23
	Synechococcus sp. JA-2-3B'a(2-13)	Ss JA-2-3B	58.5	+ 1.24	1.03	- 0.74	1.07	-- 0.55	0.98	1.17	1.11	- 0.72	1.09
	cyanobacterium UCYN-A	C_UCYN_A	31.1	1.11	0.91	0.83	1.13	0.9	1.02	0.97	1.13	-- 0.65	1.21
Gloeobacterales	*Gloeobacter violaceus PCC 7421*	Gv_PCC7421	62	+ 1.28	1.07	0.85	0.93	-- 0.57	1.04	1.05	0.93	1.08	1.13
Nostocales	*'Nostoc azollae' 0708*	Na_0708	38.33	1.13	0.92	0.89	1.03	0.84	0.95	1.09	1.17	-- 0.64	1.08
	Nostoc punctiforme PCC 73102	Np_PCC73102	41.34	1.17	0.92	0.86	1.02	0.81	0.94	1.08	1.05	0.81	+ 1.24

	Nostoc sp. PCC 7120	Ns_PCC7120	41.22	1.15	0.93	0.89	0.99	0.84	0.92	1.08	1.11	- 0.78	1.16
	Anabaena variabilis ATCC 29413	Av_ATCC29413	41.39	1.15	0.93	0.89	0.99	0.85	0.92	1.09	1.11	- 0.78	1.16
Oscillatoriales	Arthrospira platensis NIES-39	Te_IMS101	34.1	1.14	0.89	0.85	1.1	0.88	0.97	1	+ 1.24	-- 0.58	1.12
	Trichodesmium erythraeum IMS101	Ap_NIES39	44.3	1.16	1.02	0.84	0.93	0.86	0.98	0.99	+ 1.27	0.84	0.97
Prochlorales	*Prochlorococcus marinus str. AS9601*	Pm_AS9601	31.3	1.18	0.91	- 0.72	1.08	0.77	1.08	1.02	+ 1.26	-- 0.51	1.18
	Prochlorococcus marinus subsp. marinus str. CCMP1375	Pm_CCMP1375	36.4	1.15	0.92	- 0.73	1.14	0.78	1.06	1.07	1.1	-- 0.52	+ 1.27
	Prochlorococcus marinus subsp. pastoris str. CCMP1986	Pm_CCMP1986	30.8	1.17	0.92	- 0.72	1.09	0.79	1.08	1	+ 1.28	-- 0.51	1.17
	Prochlorococcus marinus str. MIT 9211	Pm_MIT9211	38	1.15	0.91	- 0.74	1.16	0.78	1.05	1.07	1.11	-- 0.52	+ 1.23
	Prochlorococcus marinus str. MIT 9215	Pm_MIT9215	31.1	1.18	0.91	- 0.71	1.08	0.77	1.09	1.02	+ 1.26	-- 0.51	1.19
	Prochlorococcus marinus str. MIT 9301	Pm_MIT9301	31.3	1.18	0.91	- 0.72	1.08	0.77	1.08	1.02	+ 1.25	-- 0.51	1.18
	Prochlorococcus marinus str. MIT 9303	Pm_MIT9303	50	1.14	1.01	- 0.78	1.07	-- 0.6	1.04	+ 1.23	0.96	- 0.74	1.22
	Prochlorococcus marinus str. MIT 9312	Pm_MIT9312	31.2	1.17	0.92	- 0.71	1.09	0.78	1.08	1.02	+ 1.26	--- 0.49	1.19
	Prochlorococcus marinus str. MIT 9313	Pm_MIT9313	50.7	1.14	1	- 0.78	1.08	-- 0.57	1.04	+ 1.24	0.95	- 0.74	1.22

ISSN 2348-6201

	Prochlorococcus marinus str. MIT 9515	Pm_MIT9515	30.8	1.17	0.92	- 0.72	1.08	0.79	1.09	1.01	+ 1.28	-- 0.52	1.16
	Prochlorococcus marinus str. NATL1A	Pm_NATL1A	35	1.16	0.92	- 0.74	1.11	0.76	1.09	1.05	1.13	-- 0.57	1.17
	Prochlorococcus marinus str. NATL2A	Pm_NATL2A	35.1	1.17	0.92	- 0.74	1.11	0.75	1.09	1.06	1.13	-- 0.57	1.18

Relation between DRDA and GC content

Mean and standard deviation was computed for GC content of genomes and each type of the dinucleotide relative abundance value for all of the 41 cyanobacteria under consideration (Table 2). From the table, it is evident that TA, CG, AC + GT and AT are least occupied (Table 2) whereas rest of the dinucleotides occupied values which are quite similar to their mean and also did not shows much deviation in their distribution pattern (Table 2).

We further carried out correlation analysis between GC content of each genome and all the possible dinucleotide combination to assess the nature of relationship shared in between them (Table 2). GC content was found to be negatively correlated with TA, CC + GG, AG + CT suggesting that GC-rich organisms are devoid of these particular nucleotides, whereas it is positively correlated with AT, TG+CA and CG, suggesting their dominance in organisms with high GC-content (Table 3).

Table 2. Mean and standard deviation of GC-percentage and dinucleotide frequency in 41 cyanobacteria.

Feature	Mean	Std. Deviation
GC %	44.5766	10.24641
AA + TT	1.1837	.05147
AT	.9800	.07029
AC + GT	.7895	.05239
AG + CT	1.0256	.07389
TA	.6944	.15981
TC + GA	1.0244	.06237
TG + CA	1.0822	.10846
CC + GG	1.1254	.13679
CG	.7490	.16061
GC	1.1290	.08663

Table 3. Correlation between GC-percentage and each of the dinucleotides in 41 cyanobacteria (**Correlation is significant at the 0.01 level (2-tailed), *Correlation is significant at the 0.05 level (2-tailed)).

	GC %	AA + TT	AT	AC + GT	AG + CT	TA	TC + GA	TG + CA	CC + GG	CG	GC
GC %	1	.218	.919**	.238	-.361*	-.839**	-.128	.735**	-.732**	.696**	-.046
AA + TT	.218	1	.106	-.183	-.591**	.003	-.383*	-.240	.272	.343*	-.360*
AT	.919**	.106	1	.181	-.405**	-.876**	.040	.740**	-.669**	.619**	-.079
AC + GT	.238	-.183	.181	1	-.490**	.036	-.432**	.064	-.209	.536**	-.331*
AG + CT	-.361*	-.591**	-.405**	-.490**	1	.094	.413**	.089	-.117	-.703**	.670**
TA	-.839**	.003	-.876**	.036	.094	1	-.280	-.859**	.775**	-.413**	-.185
TC + GA	-.128	-.383*	.040	-.432**	.413**	-.280	1	.041	-.214	-.169	.215

TG + CA	.735**	-.240	.740**	.064	.089	-.859**	.041	1	-.806**	.164	.406**
CC + GG	-.732**	.272	-.669**	-.209	-.117	.775**	-.214	-.806**	1	-.433**	-.501**
CG	.696**	.343*	.619**	.536**	-.703**	-.413**	-.169	.164	-.433**	1	-.389*
GC	-.046	-.360*	-.079	-.331*	.670**	-.185	.215	.406**	-.501**	-.389*	1

δ* differences among cyanobacterial genome

If we consider δ* differences among various members of cyanobacteria, it is evident that within species, similar result is observed as compared to between species (Figure 2). In Order Chrococcales, all members of Cyanothece species showed close similarity with each other but quite divergence was observed when Cyanothece species were compared with members of Synechococcus species and Synechocystis and Thermosynechococcus. All the members of Synechococcus showed differences with each other. However, Ss_RCC307, Ss_WH7803, Ss_WH8102 showed divergence not only with members of Cyanothece species but also with the other members of Synechococcus species and reflected a divergence across the members of this particular taxonomic unit. Members of Nostocales showed similarity in the pattern of δ* differences among them. Few members of Chrococcales i.e. C_UCYN_A, Te_BP1, Am_MBIC11017, Cs_ATCC51142, Cs_PCC7425, Cs_PCC7822 showed similar pattern of δ* differences with members of order Nostocales rather than with other members of order Chrococcales. Both the orders, Nostocales and Oscillatoriales showed similarity with each other while Nostacales was quite divergent with the order Prochlorales. Uniformity was observed when Order Prochlorales was considered with exception of Pm_MIT9303, Pm_MIT9313 which showed totally contrary pattern with rest of the members.

ISSN 2348-6201

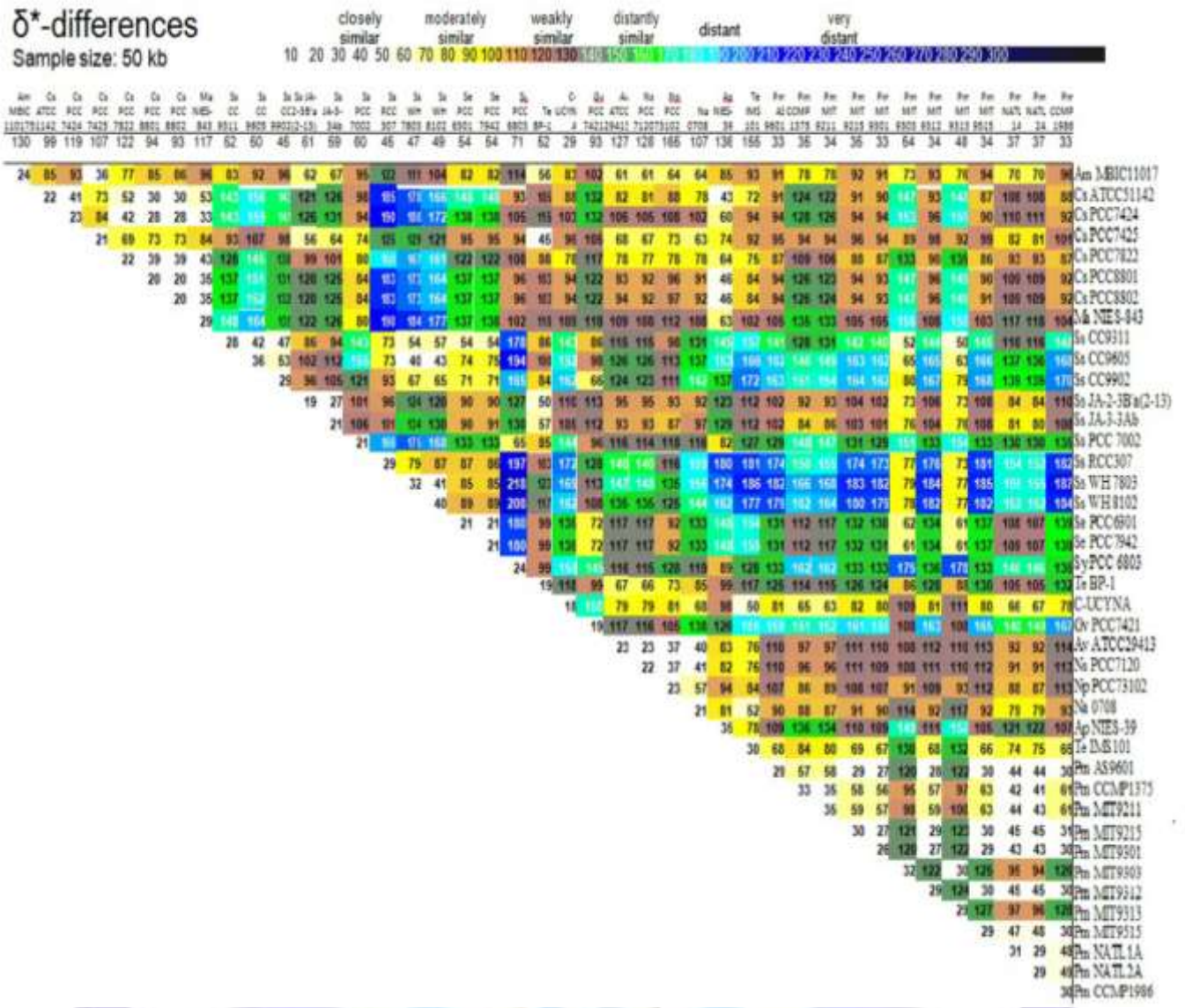

Figure 2. δ*- differences of 41 complete genomes (size ~1.4 Mb-9.0 Mb). Each genome was firstly divided into non-overlapping segments of ~50,000 bp, and then δ* value was computed for each pair of segments from the two genomes, and final result includes the average of all comparisons between the 50 kb segments multiplied by 1000 for convenience .

For clear visualization of the pattern of δ*-differences across cyanobacteria, a cluster tree based on the differences was analyzed. This has resulted in different grouping across all the members. This tree clearly divides all the cyanobacteria in two groups, one group including members of Synechococcus species, Synechocystis species, Gloeobacteria along with two members of Order Prochlorales (Figure 3). These 11 members in the first clade are those that showed distinct pattern of δ*-differences as compared to the rest members of dataset (Figure 2). Second clade include rest of the cyanobacterial species and was further subdivided into diverse clades. First sub-division separates members of Chrococcales as a separate group from the rest species. Second clade showed multiple branching patterns and included all the Prochlorales species (exception Pm_MIT9303, Pm_MIT9313) with Te_IMS101, C_UCYN_A in one branch while rest species of this clade that included members of Nostocles and Chrococcales were grouped as another branch. Among all the taxonomic orders considered only members of the Order Nostocales and Prochlorales (exception Pm_MIT9303, Pm_MIT9313) were present in a single branch while rest of the cyanobacteria were clustered in dispersed manner irrespective of their taxonomy (Figure 3).

ISSN 2348-6201

A major feature observed in the grouping of cyanobacteria was their GC content. It was observed that members with similar GC content showed similar pattern of genomic signature and grouped together as a single clade with other members which although are from different taxonomic order but have similar GC content (Figure 3).

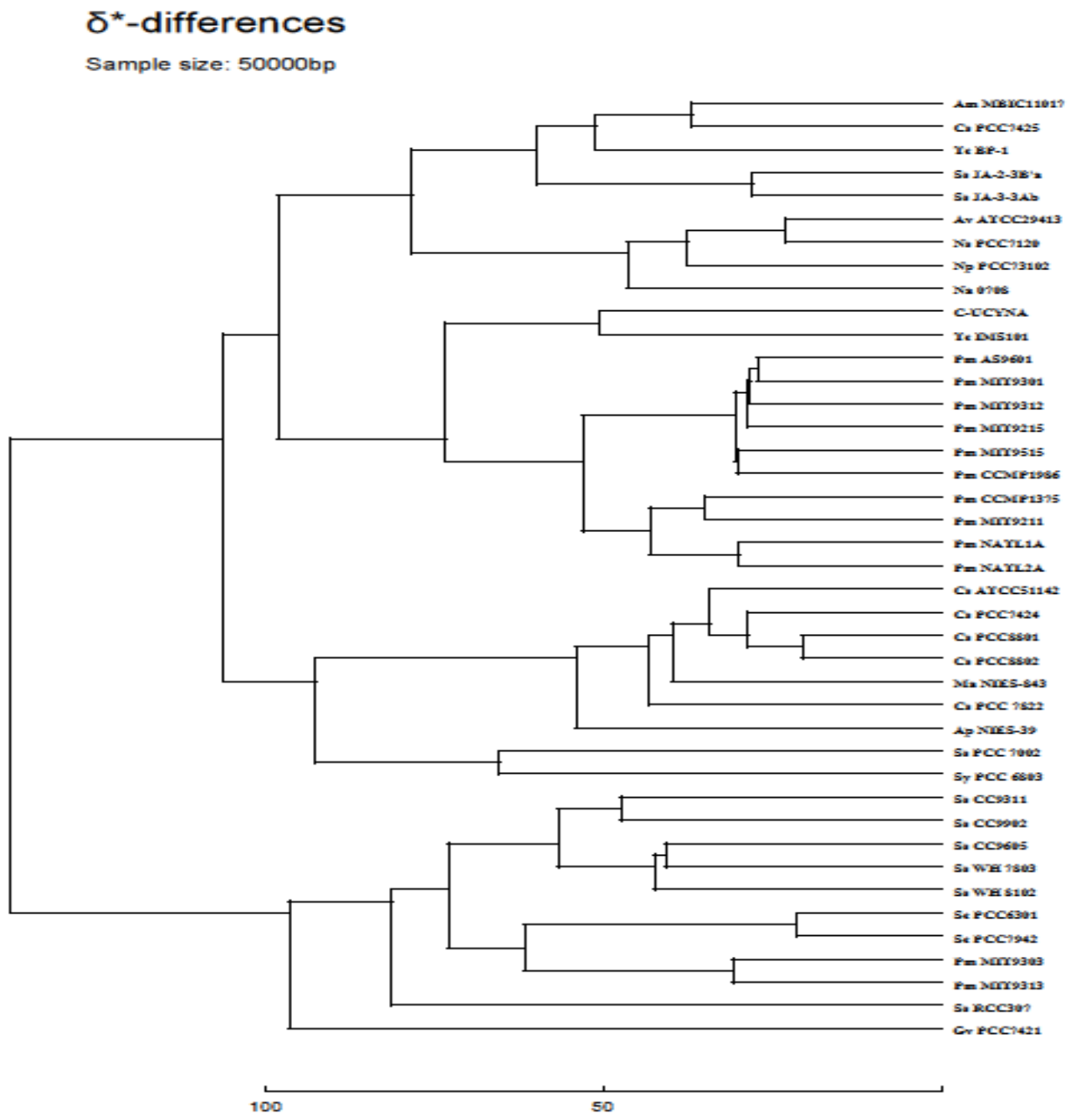

Figure 3. Clustering of cyanobacterial species on the basis of δ*-differences.

Discussion

In our analysis, TA is broadly underrepresented followed by CG and then AC+GT, AA+TT and GC occupied a major portion of dinucleotide distribution across all cyanobacterial genomes. Underrepresentation of TA seems to be influenced by GC-content of the members as it tends to decrease when there is an increase in GC content. Furthermore, GC content is negatively correlated with TA, CC + GG and AG + CT suggesting that GC-rich organisms are devoid of these particular nucleotides. GC-content is positively correlated with AT, TG+CA and CG and thus suggests their dominance in organisms with high GC-content. An interesting feature in this group of cyanobacteria is their GC content. Members with similar GC content possess similar pattern of genomic signature and grouped together as a single clade with other species that although come from different taxonomic orders, have similar GC content. Habitats also seem to influence the dinucleotide relative abundance values of the organisms because it is suggested that marine organisms show almost similar pattern of genome signature and group together as a single clade in cluster obtained on the basis of genomic signature difference. Similar is the case with organisms exhibiting either freshwater, land or multiple habitats. Average dinucleotide relative abundance distances are larger between genomes of different species in comparison to within genomes. This discrimination clearly specifies that the compositional variation of any particular genome is governed by the factors that are specific from genome to genome. Furthermore, all of the 16 dinucleotides or 10 symmetrised dinucleotides exhibit their own DNA structural preferences [4]. The dinucleotide TA remains mostly underrepresented [4, 23, 24]. It is most likely due to the lowest stacking energy of TA among all the dinucleotides which eventually allow necessary flexibility for unwinding of the DNA double helix. TA is also a part of many regulatory sequences (e.g. TATA box, polyadenylation signals) and so restricted TA usage may help to avoid improper binding of regulatory factors [1, 4]. Thus, universal under-representation of TA is an expected outcome of the extraordinarily low stacking energy in cyanobacterial genomes.

ACKNOWLEDGMENTS

Authors are thankful to Indian Council of Agricultural Research, India for providing financial support in the form of National Agricultural Bioinformatics Grid (NABG) project.

REFERENCES

[1] Karlin, S., and Burge, C. 1995. Dinucleotide relative abundance extremes: a genomic signature. Trends Genet. 11(7), 283-290.

[2] Karlin, S. 1998. Global dinucleotide signatures and analysis of genomic heterogeneity. Curr Opin Microbiol. 1(5), 598-610.

[3] Prasad, B.V.L.S., and Vemuri, M.C. 1998. Genome analysis for nucleotide interactions in fully sequenced genomes of selective prokaryotes. J. Biosci., 23(3), 255-263.

[4] Karlin, S., Ladunga, I., and Blaisdell, B.E. 1994. Heterogeneity of genomes: measures and values. Proc Natl Acad Sci USA. 91(26), 12837-12841.

[5] Jernigan, R.W., Baran, R.H. 2002. Pervasive properties of the genomic signature. BMC Genomics. 3(1), 23.

[6] Blaisdell, B.E. 1989. Effectiveness of measures requiring and not requiring prior sequence alignment for estimating the dissimilarity of natural sequences. J. Molec. Evol. 29, 526–537.

[7] Pevzner, P.A. 1992. Statistical distance between texts and filtration methods in sequence comparison. ABIOS. 8, 121–127.

[8] Mrázek, J., and Karlin, S. 1998. Strand compositional asymmetry in bacterial and large viral genomes. Proc Natl Acad Sci USA. 95, 3720–3725.

[9] Karlin, S., Mrázek, J., and Campbell, A.M. 1997. Compositional biases of bacterial genomes and evolutionary implications. J Bacteriology. 179, 3899–3913.

[10] Karlin, S., Campbell, A.M., and Mrazek, J. 1998. Comparative DNA analysis across diverse genomes. Annu. Rev. Genet. 32, 185–225.

[11] Karlin, S. 2001. Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. Trends Microbiol. 9, 335–343.

[12] van Passel, M.W., Kuramae, E.E., Luyf, A.C., Bart, A., and Boekhout, T. 2006. The reach of the genome signature in prokaryotes. BMC Evol. Biol. 6, 84.

[13] Muto, A., and Osawa, S. 1987. The guanine and cytosine content of genomic DNA and bacterial-evolution. Proc. Natl. Acad. Sci. 84, 166–169.

[14] Gelfand, M.S., and Koonin, E.V. 1997. Avoidance of palindromic words in bacterial and archaeal genome: A close connection with restriction enzymes. Nucleic Acids Res. 25, 2430–2439.

[15] Rocha, E.P.C., Viari, A., and Danchin, A. 1998. Oligonucleotide bias in Bacillus subtilis: General trends and taxonomic comparisons. Nucleic Acids Res. 26, 2971–2980.

[16] Grantham, R., Gautier, C., Guoy, M., Jacobzone, M., and Mercier, R. 1981. Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res. 9, R43–R74.

[17] Grosjean, H., and Freirs, W. 1982. Preferential codon usage in prokaryotic genes: The optimal codon-anti-codon interaction energy and the selective codon usage in efficiently expressed genes. Gene 18, 199–209.

[18] Sharp, P.M., Stenico, M., Peden, J.F., and Lloyd, A.T. 1993. Codon usage: Mutational bias, translational selection, or both? Biochem. Soc. Trans. 21, 835–841.

[19] Pride, D.T., Meinersmann, R.J., Wassenaar, T.M., and Blaser, M.J. 2003. Tetranucleotide frequency biases. Genome Res. 13, 145-158.

[20] Bohlin, J., Skjerve, E., and Ussery, D.W. 2008. Investigations of oligonucleotide usage variance within and between prokaryotes. PLoS Comput. Biol. 4(4), e1000057.

[21] Larsson, J., Nylander, J.A., and Bergman, B. 2011. Genome fluctuations in cyanobacteria reflect evolutionary, developmental and adaptive traits. BMC Evol Biol. 11, 187.

[22] Coenye, T., and Vandamme, P. 2004. Use of the genomic signature in bacterial classification and identification. System. Appl. Microbiol. 27, 175–185.

[23] Nussinov, R. 1987. Theoretical molecular biology: prospectives and perspectives. J Theor Biol. 125(2), 219–235.

[24] Burge, C., Campbell, A.M., and Karlin, S. 1992. Over- and under-representation of short oligonucleotides in DNA sequences. Proc Natl Acad Sci USA. 89(4), 1358–1362.

ISSN 2348-6201

Author' biography with Photo

Ratna Prabha

Ratna Prabha post-graduated in Bioinformatics from Banasthali University, Rajasthan, India. She is pursuing her Ph.D. from Mewar University, Rajasthan in Bioinformatics on "Whole genome approach to study cyanobacterial evolutionary diversification and adaptations to environments". She is associated with 'National Agricultural Bioinformatics Grid (NABG)' project of Indian Council of Agricultural Research (ICAR), India. Ms Prabha has been engaged in developing various bioinformtics databases and published 15 research papers in the journals of International repute and 04 book chapters. Her current research interest lies in database development, comparative microbial genome analysis, phylogenomics and pangenome analysis of prokaryotic genomes especially cyanobacteria. She has completed several bioinformatics demonstration tasks at National training programs.

Dhananjaya P. Singh

Dhananjaya P. Singh is a Senior Scientist (Biotechnology) with National Bureau of Agriculturally Important Microorganisms, Indian Council of Agricultural Research, India at Maunath Bhanjan, India. He did his Masters degree from G.B. Pant University of Agriculture & Technology, Pantnagar, India and Ph. D. in Biotechnology from Banaras Hindu University, India. His research interests lies in bioprospecting of metabolites, microbe-mediated stress management in plants, metabolomics-driven search for small molecules and bioinformatics. He is running National Agricultural Bioinformatics Grid (microbial domain) of ICAR. He has overall 110 publications including 60 research papers, two edited books, 17 book chapters, 20 popular articles, 15 reviews and one Indian Patent to his credit..