Genome Mining of Cyanobacteria Species: Diversity and Evolution of Metabolic and Biosynthetic Gene Clusters

Md. Zahid Hasan Sojib[1], Nur Ahmed Sifat[1], Md. Sarowar Hossain[2], Khondoker Moazzem Hossain[1], Md. Mahbubur Rahman[3], Md. Nazmul Hasan[1] Md. Mahmudul Islam[1]*

[1]Department of Genetic Engineering and Biotechnology, Daffodil International University, Dhaka, Bangladesh
[2]Computational Biology Research laboratory, Department of Pharmacy, Daffodil International University, Dhaka, Bangladesh
[3]Department of Nutrition and Food Engineering, Daffodil International University, Dhaka, Bangladesh
*Correspondence; Email: mahmudul.geb@diu.edu.bd

Abstract

Cyanobacteria are well recognized for producing a large number of secondary metabolites, many of which are employed extensively in agriculture, animal health, and human medicine. Understanding the evolutionary connections of *Cyanobacteria* species, genome-wide variety, and distribution patterns of BGCs is essential for directing the efficient prioritization of certain biosynthetic gene clusters (BGCs) for drug development and focusing on the most prolific producer strains. We examined the diversity of key groups of BGCs in over 8,000 publically accessible *Cyanobacteria* genomes using phylogenetic and genomic techniques. Even among closely related strains, the *Cyanobacteria* phylogeny exhibits diverse distribution patterns and a large variety of BGCs, according to genome mining. NRPS, nrps-like, t1pks, terpene, and terpene-precursor are the most prevalent BGCs. Additionally, we discovered that several species of *Cyanobacteria* have BGCs that are known to encode antimicrobial substances. We found that the BGCs carried by strains that are thought to belong to the same species might differ significantly, indicating that strain-level genome sequencing can reveal high levels of BGC variation and potentially beneficial derivatives of any one chemical. These results imply that a strain-level technique for investigating secondary metabolites for therapeutic applications offers a different or supplementary method for identifying new pharmacological chemicals from microorganisms..

Keywords: *Cyanobacteria,* Secondary metabolites, Genomic diversity, Biosynthetic gene clusters (BGCs), Genome mining, Natural products, genome comparison

Introduction

Members of the phylum Bacteria Among the oldest and most metabolically diverse organisms on Earth, cyanobacteria are essential to primary production, oxygenic photosynthesis, and global biogeochemical cycles [Saleem et al., 2025]. They live in a variety of settings, such as hot springs and deserts, as well as freshwater and marine environments. In addition to their ecological significance, cyanobacteria are abundant producers of secondary metabolites that have uses in biotechnology, agriculture, and medicine [Singh et al., 2027]. Microcystins, nodularins, and cyanobactins are examples of bioactive substances that show promise as important sources of new medications and agricultural chemicals. The quest for natural compounds with distinct chemical structures and modes of action has accelerated because to the growing danger of multidrug-resistant infections and the continuous demand for novel treatment medicines [Ahmed et al., 2024]. However, conventional techniques of discovery often result in the repetitive isolation of known molecules, which slows innovation. This problem has been solved by developments in high-throughput genome sequencing and bioinformatics, which make it possible to use genome mining techniques that uncover microbes' latent metabolic potential [Lema et al., 2023]. These methods concentrate on finding biosynthetic gene clusters (BGCs), which encode pathways for generating a variety of secondary metabolites, including terpenes, polyketides, nonribosomal peptides, and ribosomally synthesized and post-translationally modified peptides (RiPPs). The amazing variety of BGCs found in cyanobacterial genomes suggests that their complete metabolic potential is still largely unknown, since many of these genes are either quiet or weakly expressed in lab settings [Cameron et al., 2024]. Because of evolutionary processes including horizontal gene transfer, gene duplication, and ecological adaption, the distribution of these clusters differs across lineages. Recent genome mining research raises significant issues about BGC diversity and evolution while highlighting the enormous and mostly unrealized potential of cyanobacteria as sources of new bioactive chemicals [Malit et al., 2022; Krishan et al., 2026].

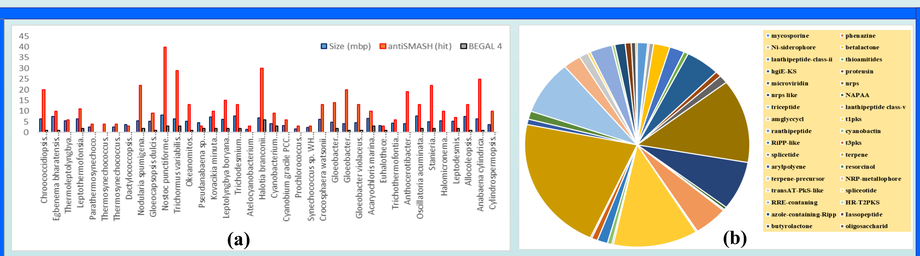

Figure 2. The distribution of key types of BGCs in Cyanobacteria species and the relationship between genome size and the number of BGCs on each genome. **(a)** The number of BGCs extracted from each genome using various genome mining techniques is contrasted with the genome's size. **(b) AntiSMASH** hit distribution for the main BGC classes.

Table1.Reference genomes of *Cyanobacteria* species studied with different hits from different genome mining tools for BGCs.

species	Isolation source	Size (mbp)	Genes	GC content(%)	antiSMASH (hit)	BEGAL 4	KS Domain	C Domain
Chroococcidiopsis thermalis PCC 7203	soil	6.31	6242	44.5	20	1	5	27
Egbenema bharatensis FJBBOT2019	biocrust	7.5	5960	50.5	10	1	4	8
Thermoleptolyngbya sichuanensis A183	mat	5.52	4623	56.5	6	0	4	0
Leptothermofonsia sichuanensis E412	water	6.42	5620	51	11	2	2	1
Parathermosynechococcus lividus PCC 6715	N/A	2.65	2596	53.5	4	0	2	0
Thermosynechococcus elongatus BP-1 DNA	N/A	2.59	2547	54	4	0	2	0
Thermosynechococcus sichuanensis E542	N/A	2.65	2596	53.5	4	0	2	0
Dactylococcopsis salina PCC 8305	solar lake	3.78	3822	42.5	3	0	2	1
Nodularia spumigena CCY9414	water	5.46	4842	41	22	2	15	31
Gloeocapsopsis dulcis strain AAB1	desert	5.29	5465	42.5	9	1	4	3
Nostoc punctiforme PCC 73102	N/A	8.23	7729	41.5	40	3	34	43
Trichormus variabilis ATCC 29413	N/A	6.36	5993	41.5	29	3	15	15
Okeanomitos corallinicola TIOX110	coral reef	5.14	4723	37	13	1	6	22
Pseudanabaena sp. PCC 7367		4.55	3992	46.5	3	2	2	0
Kovacikia minuta CCNUW1	fresh water	7.29	8301	49.5	10	2	2	1
Leptolyngbya boryana IAM M-101	N/A	6.17	6446	47	15	2	2	4
Trichodesmium erythraeum IMS101	N/A	7.75	5449	34	13	2	5	2
Atelocyanobacterium thalassa	flow cytometry	1.44	1224	31	3	0	1	0
Halotia branconii CENA392	whale bone	6.86	6323	40.5	30	6	13	24
Cyanobacterium aponinum PCC 0605	fresh water	4.11	3601	35	9	3	3	0
Cyanobium gracile PCC 6307	lake water	3.34	3427	68.5	6	0	1	0
Prochlorococcus marinus CCMP1375	N/A	1.75	1911	36.5	3	0	1	0
Synechococcus sp. WH 8102	N/A	2.43	2711	59.5	3	0	2	0
Crocosphaera watsonii WH 8501	czechia trebon	6.13	6198	37	13	0	6	24
Gloeobacter kilaueensis JS1	cave	4.72	4552	60.5	14	2	8	7
Gloeobacter morelensis MG652769	cave	4.12	4974	62	20	1	22	10
Gloeobacter violaceus PCC 7421	N/A	4.65	4603	62	13	1	16	0
Acaryochloris marina MBIC11017	N/A	6.5	7652	47.5	10	3	2	0
Euhalothece natronophila Z-M001	soda lake	3.32	3543	41	3	1	2	0
Trichothermofontia sichuanensis B231	water	4.43	3800	54	6	0	1	1
Anthocerotibacter panamensis C109	plant leaf surface	4.18	4015	55.5	19	0	15	10
Oscillatoria acuminata PCC 6304	soil	7.68	6151	47.5	13	2	4	11
Stanieria cyanosphaera PCC 7437	botanical garden	5.04	5049	36	22	2	7	8
Halomicronema hongdechloris C2206	somatolites grown	5.57	5233	54.5	10	0	7	7
Leptodesmis sichuanensis A121	water	5.34	5003	50.5	7	2	2	1
Allocoleopsis franciscana PCC 7113	soil	7.47	6797	46	13	1	4	10
Anabaena cylindrica PCC 7122	pond water	6.39	6196	39	25	1	11	13
Cylindrospermopsis raciborskii Cr2010	fresh water	3.75	3416	40	10	0	7	8

Table 2. List of different *Nostoc punctiforme* reference genomes with different hit.

Species Name	source	Size (Mbp)	Gene	GC content(%)	antiSMASH (hi	BAGEL	KS Domai	C Domain
Nostoc flagelliforme CCNUN1	desert soil	8.36	9790	42	29	3	10	28
Nostoc sp. FACHB-892	N/A	7.02	7016	41.5	21	3	9	18
Nostoc sphaeroides CCNUC1	rice field	6.48	8672	41.5	12	0	5	1
Nostoc sp. C110	N/A	7.17	7168	41.5	29	3	22	23
Nostoc cf. commune SO-36	N/A	6.8	6847	41.5	22	0	8	21
Nostoc sp. DSM 114167	coralloid root	8.47	7119	41.5	34	2	23	33
Nostoc sp. UHCC 0302	lake sediment	8.38	8492	41	32	3	16	22
Nostoc sp. DSM 114161	coralloid root	8.68	7247	41.5	47	6	12	22
Nostoc sp. TCL26-01	N/A	6.76	6425	41	24	3	9	
Nostoc sp. PCC 7120	N/A	6.41	6169	41.5	23	3	13	14
Nostoc sp. MS1	river water	7.14	7043	40.5	19	2	6	5
Nostoc sp. NIES-3756	N/A	6.46	6092	40.5	22	2	6	8
Nostoc sp. ATCC 53789	lichen thallus	7.34	7338	41.5	32	3	22	29
Nostoc sp. C117	N/A	7.88	6931	41	48	4	26	26
Nostoc sp. TCL240-02	N/A	7.88	6907	41.5	35	3	24	22
Nostoc sp. C057	N/A	7.49	8158	41.5	31	5	22	29
Nostoc piscinale CENA21	solimoes river	7.09	6478	40.5	20	3	10	16
Nostoc sp. PCC 7524	moderate hot spring	6.63	5674	41.5	22	4	11	1
Nostoc sp. CENA543	saline-alkaline lake	6.99	6149	41	38	6	11	22
Nostoc sp. PCC 7107	water from shallw pond	6.32	5531	40.5	16	2	9	7
Nostoc edaphicum CCNP1411	bay water	7.73	7073	41.5	35	2	19	46
Nostoc azollae 0708	N/A	5.35	5697	38.5	12	1	7	2
Nostoc sp. UHCC 0926	lichen surface	6.61	7355	42.5	15	0	6	3
Nostoc sp. UHCC 0870	brackish water	6.42	6197	41	31	5	12	27
Nostoc sphaeroides strain Kutzing En	water	6.53	6108	41.5	12	0	5	1
Nostoc sp. C052	N/A	7.82	8527	41.5	37	5	12	41

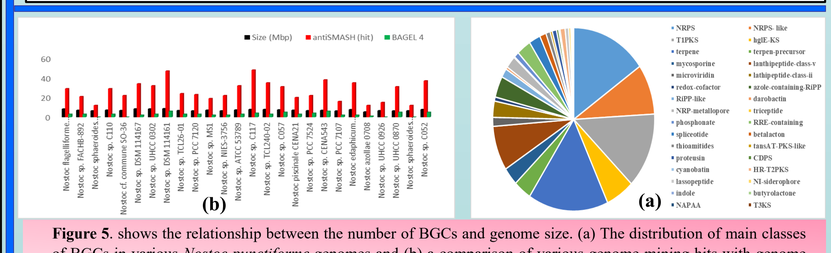

Figure 5. shows the relationship between the number of BGCs and genome size. (a) The distribution of main classes of BGCs in various *Nostoc punctiforme* genomes and (b) a comparison of various genome mining hits with genome size (Mbp) in *Nostoc punctiforme* subspecies.

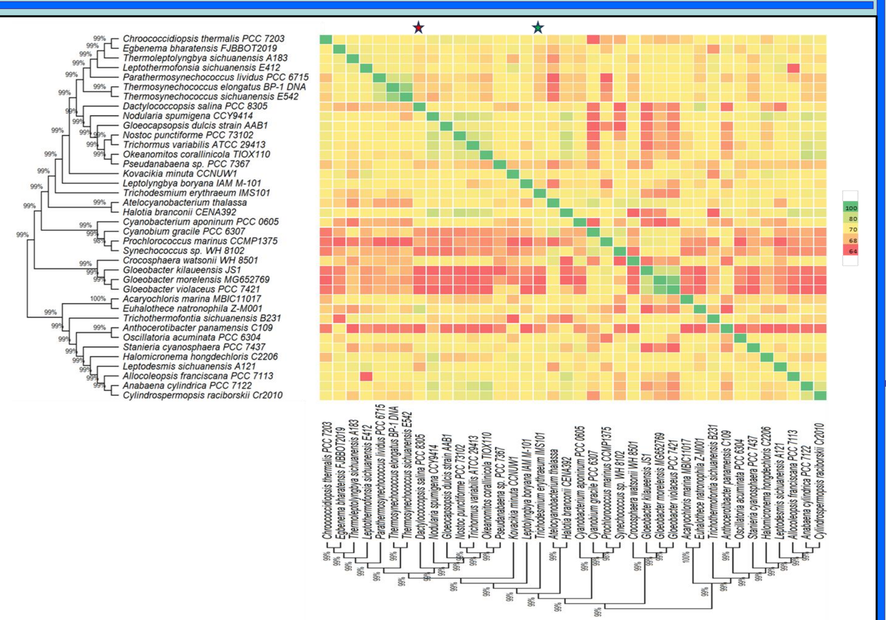

Figure 3: *Cyanobacteria* species' whole genomes are similar. The evolutionary tree in **Figure 1** and the comparison follow the same order. A line that extends from the top left to the bottom right corners of the genome is used for all comparisons between a genome and itself. The number of genes that are similar between two genomes serves as the numerator in each comparison, while the genome that each column represents serves as the denominator. A * (green) indicates the smallest genome, while a * (red) indicates the largest genome.

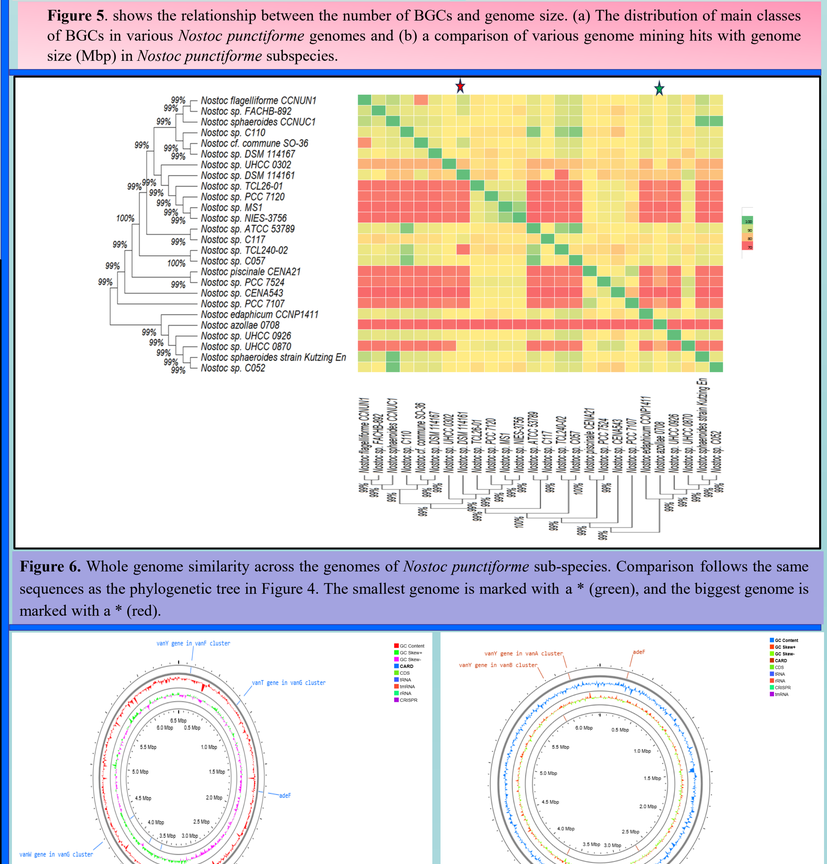

Figure 4: *Nostoc punctiforme* subspecies phylogenetic tree with antiSMASH-determined gene numbers, isolation sources, and NP BGC numbers. 16s rRNA sequences taken from the genome using the greatest likelihood approach are used to construct the phylogenetic tree. Two bar graphs display the number of BGCs on the right and the genome size in Mbp of genes on the left, colored by habitats. The evolutionary tree's order is maintained by the species in these two bar plots. Separate hybrid clusters are shown. Below the bar-plots are the colors that correspond to the 32 primary NP classifications and habitat types. N/A: Not accessible.

Figure 6. Whole genome similarity across the genomes of *Nostoc punctiforme* sub-species. Comparison follows the same sequences as the phylogenetic tree in Figure 4. The smallest genome is marked with a * (green), and the biggest genome is marked with a * (red).

Methodology

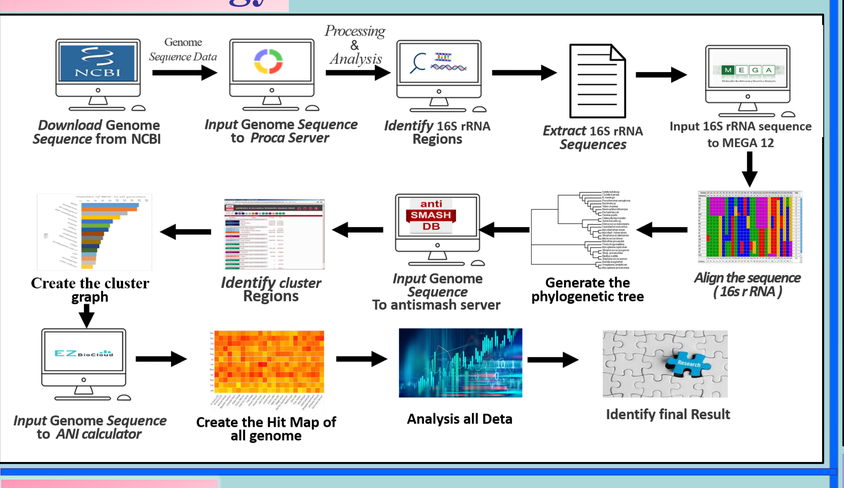

Figure 8. (a) The identification of the most common resistant gene in Cyanobacteria species **(b)** The identification of the most common resistant gene in Cyanobacteria sub-species.

Results

1.Distribution and Diversity of Biosynthetic Potential in Cyanobacteria at Species Level

From the NCBI, selected (38) cyanobacterial genomes with the *16s rRNA* gene were analyzed for biosynthetic potential, while others were excluded from phylogenetic analysis.

1.1Putative BGC Prediction by antiSMASH in Cyanobacteria Species Genomes

Using antiSMASH, diverse biosynthetic gene clusters (BGCs) were identified, showing high secondary metabolite diversity. The diversity of these species influences the phylogenetic diversity and hetero geneity **(Figure 1).**

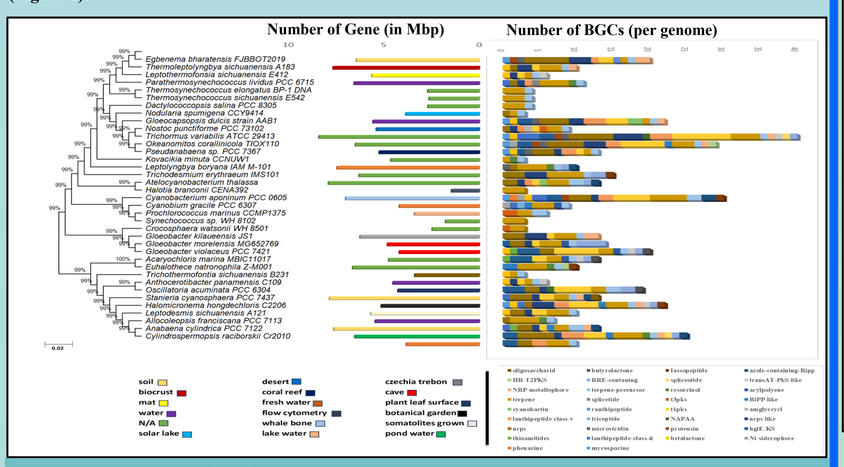

Figure 1. Phylogenetic tree of *Cyanobacteria* along with the gene numbers, isolation sources, and NP BGCs number determined by antiSMASH. The phylogenetic tree is built using 16s rRNA sequences extracted from the genomes based on the maximumlikelihood method. Two bar-plots show genome size in Mbp of genes on the left, colored by habitats, and the number of BGCs on the right. Species in these two bar-plots keep the same order as the phylogenetic tree. Hybrid clusters are shown separately. The colors matching to habitat types and 26 major NP classes are displayed below the bar-plots. **N/A: Not Available.**

Discussion

- Analyzed around 8000 *Cyanobacteria* genomes, revealing extensive diversity in biosynthetic gene clusters (BGCs) and core gene clusters (CGCs).
- BGC distribution varies widely, even among closely related strains, with hybrid BGCs increasing diversity.
- Significant inter-strain variation shows that strains within the same species can have distinct metabolomic profiles.
- Hybrid BGCs and inter-strain variation significantly expand biosynthetic potential.
- Hybridization of BGCs contributes to novel metabolite production.
- Future work should use deep sequencing and genome-wide analyses to better understand BGC diversity and drug discovery potential.

References

1.Saleem A, Anwar S, Saud S, Kamal T, Fahad S, Nawaz T. Cyanobacteria diversity and ecological roles: insights into cyanobacterial adaptations and environmental implications. Journal of Umm Al-Qura University for Applied Sciences. 2025 Aug 27:1-9.

2.Singh R, Parihar P, Singh M, Bajguz A, Kumar J, Singh S, Singh VP, Prasad SM. Uncovering potential applications of cyanobacteria and algal metabolites in biology, agriculture and medicine: current status and future prospects. Frontiers in microbiology. 2017 Apr 25;8:515.

3.Lema NK, Gemeda MT, Woldesemayat AA. Recent advances in metagenomic approaches, applications, and challenges. Current Microbiology. 2023 Nov;80(11):347.

4.Ahmed SK, Hussein S, Qurbani K, Ibrahim RH, Fareeq A, Mahmood KA, Mohamed MG. Antimicrobial resistance: Impacts, challenges, and future prospects. Journal of Medicine, Surgery, and Public Health. 2024 Apr 1;2:100081.

5.Cameron ES, Sanchez S, Goldman N, Blaxter ML, Finn RD. Diversity and specificity of molecular functions in cyanobacterial symbionts. Scientific Reports. 2024 Aug 12;14(1):18658.

6.Krishan B, Kumar A, Azmi W. Advances in tools, strategies, and applications of mining of microbial genomes for novel antimicrobials: a comprehensive review. Folia Microbiologica. 2026 Apr 22:1-25.

7.Malit JJ, Leung HY, Qian PY. Targeted large-scale genome mining and candidate prioritization for natural product discovery. Marine drugs. 2022 Jun 16;20(6):398.

Acknowledgement

The work is done by the supported of Bioinformatics Laboratory, Department of Genetic Engineering and Biotechnology, Faculty of Health and Life Sciences, Daffodil International University, Dhaka, Bangladesh.