Scientific Research Publishing

Journal of Geoscience and Environment Protection, 2017, 5, 198-213
http://www.scirp.org/journal/gep
ISSN Online: 2327-4344
ISSN Print: 2327-4336

Current Trend of Metagenomic Data Analytics for Cyanobacteria Blooms

JianDong Huang, Huiru (Jane) Zheng*, Haiying Wang

School of Computing and Mathematics, Ulster University, Jordanstown, Northern Ireland, UK
Email: *h.zheng@ulster.ac.uk, hy.wang@ulster.ac.uk, jd.huang@ulster.ac.uk

How to cite this paper: Huang, J.D., Zheng, H.R. (J.) and Wang, H.Y. (2017) Current Trend of Metagenomic Data Analytics for Cyanobacteria Blooms. *Journal of Geoscience and Environment Protection*, 5, 198-213.
https://doi.org/10.4236/gep.2017.56018

Received: June 1, 2017
Accepted: June 27, 2017
Published: June 30, 2017

Abstract

Cyanobacterial harmful algal blooms are a major threat to freshwater ecosystems globally. To deal with this threat, researches into the cyanobacteria bloom in fresh water lakes and rivers have been carried out all over the world. This review presents an overlook of studies on cyanobacteria blooms. Conventional studies mainly focus on investigating the environmental factors influencing the blooms, with their limitation in lack of viewing the microbial community structures. Metagenomics study provides insight into the internal community structure of the cyanobacteria at the blooming, and there are researchers reported that sequence data was a better predictor than environmental factors. This further manifests the significance of the metagenomic study. However, large number of the latter appears to be confined only to present snapshoot of the microbial community diversity and structure. This type of investigation has been valuable and important, whilst an effort to integrate and coordinate the conventional approaches that largely focus on the environmental factors control, and the Metagenomics approaches that reveals the microbial community structure and diversity, implemented through machine learning techniques, for a holistic and more comprehensive insight into the cause and control of Cyanobacteria blooms, appear to be a trend and challenge of the study of this field.

Keywords

Cyanobacteria Blooms, Harmful algal, Metagenomics, Machine Learning, Environmental Factors, Next Generation Sequencing Techniques (NGS), 16S rRNA, Fresh Water Ecosystem, Lakes

1. Introduction

Cyanobacteria blooms are commonly associated with toxin production in drinking water supplies and have been a severe risk to human beings health [1]

[2]. The blooms have become a major threat to freshwater ecosystems globally and a worldwide challenge [3] [4]. To deal with the threat and challenge, studies of the cyanobacteria blooms have been carried out on the fresh water systems in Asia [5]-[14], Europe [15]-[19], North America [20]-[26], Oceania [27] [28] and Africa [29] [30]. For decades, it has seen that the study appears to be characterized mainly by investigating the nutrient control (such as Nitrogen and Phosphorus) and the influence of other environmental factors (temperature and pH, for example) on the blooms, and deploying hydrodynamic and microbial combined models to predict the blooms, or introducing machine learning methods such as Artificial Neural Network (ANN), with environmental factors as their input variables, for gaining understanding of the cause and development of the microbial community's explosive reproduction [8] [11] [31] [22] [33]. On the other hand, new generation high-throughput sequencing techniques, based on the system of 16S rRNA gene, makes it possible to quickly examine the composition of the microbial community comprehensively in different habitats, enabling insight into profiles of the community composition [34] [35] [36]. With the rapid development of next generation sequencing techniques, metagenomic data analysis has been applied to the Cyanobacteria bloom study. Applying Metagenomics to investigate the genetic and metabolic diversity of the mixed populations helps understand the interactions of different microbial populations and their functions in the blooming process. A recent research carried out by Tromas *et al.* [25] shows that sequence data was a better predictor than environmental factors. In this article, we present a review of recent study of harmful Cyanobacteria bloom, with more attention paid to the cases in China that has been suffering from severely and critically growing water quality problem [37]-[38].

2. Conventional Approach: Investigating the Relation between Environmental Factors and the Blooms

The conventional approach is focused on impact of environmental factors to the bloom. Kong, F. and Fao, G. [39] noticed the environmental elements as control factors of the algae blooms. Temperature and dissolve oxygen on the sediment surface were observed, and they concluded that the hydrological and meteorological condition would cause the algae to float up to the water surface and then form the water bloom. Wilhelm *et al.* [40] investigated the relationships between nutrients, cyanobacterial toxins and the microbial community in Taihu, China. They provide an independent confirmation that both total nitrogen and total phosphorus concentrations are strongly related to cyanobacterial biomass in the system and demonstrated that both nitrogen and phosphorus inputs play a role in microbial community biomass production and structure. They also indicated that the toxicity of the community is not closely coupled to key factors leading to bloom formation. In a work by McCarthy *et al.* [41], nitrogen dynamics and microbial food web structure during a summer cyanobacterial bloom in Lake Taihu were studies, they found that N limits or colimits primary production in and near central Lake Taihu, contrary to the previous paradigm of exclusive P limitation, and saw this result an example to show the importance of characterizing N cycling in freshwater systems, where most studies have focused on P dynamics. They also stated the importance of water column N recycling relative to sediment processes. An example from North America is a study undertook by Graham *et al.* [20]. Physicochemical data were collected from 241 lakes in Missouri, Iowa, northeastern Kansas, and southern Minnesota U.S.A., to determine the environmental variables associated with high concentrations of the cyanobacterial hepatotoxin microcystin (MC), during May-September 2000-2001. Relationships between particulate MC values and environmental variables were developed using nonparametric Spearman–Rank correlation (a = 0.05). The following environmental factors were measured: Secchi transparency, surface temperature, total phosphorus (TP), total nitrogen (TN), TN:TP ratio, and total suspended solids (TSS); chlorophyll (Chl), and Chl:TP ratio. They found that the presence and concentration of microcystin increase along a gradient of increasing lake trophic status. Ma *et al.* [42] reported the influence of N, P and pH on Microcystis growth and colony formation in field simulation experiments in Lake Taihu (China). Krausfeldt *et al.* [43] examined the spatial and temporal variability in the nitrogen cyclers of hypereutrophic Lake Taihu. These studies focused on the physical and chemical parameters to examine how the environmental and biological variables were associated with the cyanobacteria blooms. Examining environmental parameters such as water temperature, solar radiation, precipitation, water transparency, pH, DO, and nutritious elements e.g. TN, TP, DN, DP, $PO_{4\text{-}}P$, $NH_{4\text{-}}N$. $NO_{3\text{-}}N$, etc (e.g. [11]), characterized the early studies of the cyanobacteria blooms.

Hydrodynamics modelling, statistical methods and machine learning approaches have been facilitating the research [8] [31] [44] [45] [46]. Regression and multivariate analyses by principal component and classifying analysis were performed in the study of Wu *et al.* [31] on cyanobacterial toxin microcystin in 30 subtropical shallow lakes in the middle and lower reaches of the Yangtze River area in China; Li *et al.* [8] used a coupled hydrodynamic–algal biomass model for forecasting short-term cyanobacterial blooms in Lake Taihu; the model was applied to predict the occurrences of the algae blooms of the next 3 days in Lake Taihu during April to September in 2009 and 2010. They reported that independent evaluations from remote sensing images and boat survey data showed that the accuracy of these bloom forecasts was more than 80%. Yabunaka *et al.* [11] applied machine learning techniques for modelling algal bloom dynamics. Artificial Neural Network models, through genetic programming, were used to model and predict the blooms in Tolo Harbour, Hong Kong. The input variables were 10 parameters, four nutrients ($PO_{4\text{-}}P$, $NH_{4\text{-}}N$. $NO_{3\text{-}}N$, and Si), four physic-chemical conditions (water temperature. transparency, DO, and pH), plus two zooplankton species, and the output variables were the chlorophyll *a* concentration or the biomass of specific phytoplankton species. Vilán *et al.* used support vector machines and multilayer perceptron networks from cyanobacterial concentrations determined experimentally in the Trasona reservoir in Northern

Spain, to build a cyanotoxin diagnostic model [46]. They reported that the SVR (support vector regression) and MLP (multilayer perceptron) techniques predict the observed actual cyanobacteria blooms from 2006 to 2010 more effectively and accurately than traditional regression models. In this research, the output variable was cyanotoxins and the input variables were a number of biological and physical-chemical variables; the biological parameters included *Microcystis aeruginosa*, *Woronichinia naegeliana*, other cyanobacteria species, diatoms, *chrysophytes*, *chlorophytes* and other phytoplankton species; the physical-chemical variables were: water temperature, ambient temperature, secchi disk depth, turbidity, total phosphorus concentration, total nitrogen concentration, nitrate concentration, nitrite concentration, ammonium ion concentration, dissolved oxygen concentration, conductivity, alkalinity, calcium concentration, and pH.

Zhou *et al.* [12], in their study of the influence of turbulence on MCs concentrations in Lake Taihu during cyanobacterial bloom periods, investigated how toxic Microcystis and MCs production may be affected by wind-driven turbulence, using a mesocosm experiment. The study aims to deliver deeper insights into the competition of toxic *Microcystis* and MCs regulation, and understand the coupling of MCs production and turbulence. In this study, a 6-day mesocosm experiment was carried out to evaluate the effects of wind wave turbulence on the competition of toxic Microcystis and MCs production in highly eutrophicated and turbulent Lake Taihu, China. Under turbulent conditions, MCs concentrations (both total and extracellular) significantly increased and reached a maximum level 3.4 times higher than in calm water. Specifically, short term (about 3 days) turbulence favored the growth of toxic Microcystis species, allowing for the accumulation of biomass which also triggered the increase in MCs toxicity. Moreover, intense turbulence raises the shear stress and could cause cell mechanical damage or cellular lysis resulting in cell breakage and leakage of intracellular materials including the toxins. The results indicate that short term (about 3 days) turbulence is beneficial for MCs production and release, which increase the potential exposure of aquatic organisms and humans. This study suggests the importance of water turbulence in the competition of toxic Microcystis and MCs production, and provides new perspectives for control of toxin in CyanoHABs-infested lakes.

Although encouraging outcomes have obtained from the studies represented by aforementioned cases, localization of mainly examining the environmental parameter makes it still inadequate for a deep and holistic inspection of the bloom phenomenon to be gained, especially in terms of the diversity and composition or structure of the microbial community of the bloom forming cyanobacteria. Recent research has demonstrated that, in addition to the conventional methods, it is possible to use pre-bloom sequence data to predict the number of days until a bloom event occurs, with good accuracy; sequence data appears to be a strong predictor, similar or better than prediction with environmental variables [25].

3. Next Generation Sequencing Techniques Based on the System of 16S rRNA Gene for Microbial Community Profiling

Application of Metagenomics [47] has been accompanied with high speed throughput Next-Generation Sequencing (NGS) that surpass traditional Sanger approach for DNA isolation and sequencing. Development of high-throughput DNA sequencing techniques brings about the progress in microbial community profiling using 16S rRNA [34]-[36], [48] for analysing population structure of cyanobacterial blooms. High-throughput DNA sequencing techniques speed up the analysis through bypassing the need of isolation or cultivation of microorganisms [49]-[50], i.e., the cyanobacteria concerned. The next-generation sequencing technology shows advantages in its high flex, short test period, low cost and repeatability [51]-[52]. This culture-independent, molecular way of analysing environmental samples of cohabiting microbial populations has opened up fresh perspectives on microbiology [53]. **Figure 1** is a microbial community analysis pipe line diagram:

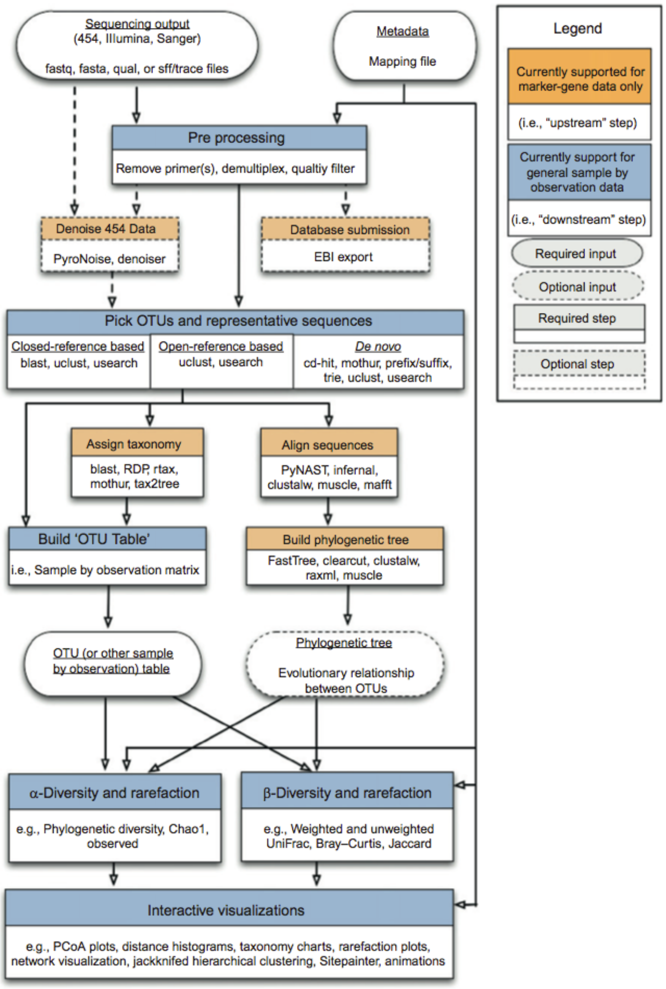
Sequencing output
(454, Illumina, Sanger)
fastq, fasta, qual, or sff/trace files
Metadata
Mapping file
Legend
Currently supported for marker-gene data only
(i.e., "upstream" step)
Currently support for general sample by observation data
(i.e., "downstream" step)
Required input
Optional input
Required step
Optional step
Pre processing
Remove primer(s), demultiplex, qualtiy filter
Denoise 454 Data
PyroNoise, denoiser
Database submission
EBI export
Pick OTUs and representative sequences
Closed-reference based
blast, uclust, usearch
Open-reference based
uclust, usearch
De novo
cd-hit, mothur, prefix/suffix, trie, uclust, usearch
Assign taxonomy
blast, RDP, rtax, mothur, tax2tree
Align sequences
PyNAST, infernal, clustalw, muscle, mafft
Build 'OTU Table'
i.e., Sample by observation matrix
Build phylogenetic tree
FastTree, clearcut, clustalw, raxml, muscle
OTU (or other sample by observation) table
Phylogenetic tree
Evolutionary relationship between OTUs
α-Diversity and rarefaction
e.g., Phylogenetic diversity, Chao1, observed
β-Diversity and rarefaction
e.g., Weighted and unweighted UniFrac, Bray–Curtis, Jaccard
Interactive visualizations
e.g., PCoA plots, distance histograms, taxonomy charts, rarefaction plots, network visualization, jackknifed hierarchical clustering, Sitepainter, animations

Figure 1. Microbial community analysis pipeline (https://sites.google.com/site/knightslabwiki/qiime-workflow).

Table 1. Some environmental parameters seen in Cyanobacteria bloom study.

Water temperature (WT)
Ambient temperature
Secchi disk depth(SD)
Transparency
Turbidity
Solar radiation
Total phosphorus (TP)
Total nitrogen (TN)
NH_4-N
NO_3-N
Ammonium ion concentration
Dissolved oxygen concentration (DO)
Conductivity
Alkalinity
Calcium concentration
Total suspended solids (TSS)
Si
pH
Salinity
Chlorophyll a

4. Applying Metagenomics to Characterize the Structure and Function of Microbial Community in Fresh Water Ecosystem

Some examples are inspected in this section, with special attention paid to lakes in China and North America for their significant influence on the large number of populations.

4.1. Microbial Community Structures

Through a metagenomic approach, Xie *et al.* investigated the relationship between *Microcystis* and the associated bacteria [10]. They analyzed cyanobacteria-dominated bloom communities from Lake Taihu, China, applying a visualization-enhanced binning method they developed. By analyzing the metabolic pathways of the microbial community, cooperative interactions among the complex species were indicated. The study revealed that while all heterotrophic bacteria were dependent upon *Microcystis* for carbon and energy, Vitamin B12 biosynthesis, which is required for growth by *Microcystis*, was accomplished in a cooperative fashion among the bacteria. The analysis also suggests that individual bacteria in the colony community contributed a complete pathway for degradation of benzoate, which is inhibitory to the cyanobacterial growth. Next-gen-

Table 2. Summary of the lakes examined in this section.

Name	Location
Taihu	East China
Poyanghu	South east China
Lakes in Nanjing City	East China
Lakes in Yunnan plateau	South west China
Lake Erie	North America
Grand Lakes	North America
Lake Champlain	North east USA, across the Canada-USA border
Lough Corrib and Ballyquirke Lough	Ireland

eration sequencing was applied in the study by Huang *et al.* for the microbial diversity in lake-river ecotone of Poyang Lake, China as well [9]. They aimed to identify the micro diversity in different lake-river ecotone, and to explore the evolution and adaptation of the microbial population to changing environmental conditions. The results showed the major Poyang Lake had the largest microbial population, followed by Yao Lake, Ganjiang River and Raohe River. Based on the Shannon and Simpson Index, major Poyang Lake had the largest biodiversity of microbial communities, followed by Ganjiang River, Yao Lake, and Raohe River. Microbial characteristics vary with the TN and TP concentration, for instance, the nitrifying bacteria were relatively rich in Yao Lake and Ganjiang River ecotone, and the polyphosphate-accumulating organisms (PAO) in Raohe River were richer than those in Ganjiang River. In a study carried out by Zhao *et al.*, high-throughput sequencing was employed to investigate the seasonal variations in the composition of bacterioplankton communities in six eutrophic urban lakes of Nanjing City, China [54]. The results showed that temperature, pH and NO_{3-}-N were the most important factors influencing the composition of the bacterioplankton community. The length and direction of temperature arrow suggested strong impact to the summer community. Temperature was orthogonal with the other two arrows (pH and N_{O3-}-N), suggesting temperature explains variation not explained by pH and NO_{3-}-N. The results demonstrated that co- occurrence in freshwater bacterioplankton communities within six urban lakes varied in different seasons. Moreover, Cyanobacteria played different roles in the ecological network of each season. In summer, Cyanobacteria were dominant which may result from the strong co-occurrence pattern, suitable temperature and eutrophication. The bacterial community within a module maintained similar ecological niches. The analysis of the relationships between the module eigengenes and environmental variables provided a highly simplified version of the complex effects of environmental variables on the bacterial communities. Module eigengene analysis indicated that temperature only affected some Cyanobacteria members, while others were mainly affected by the nitrogen associated factors. Overall, this study applied network analysis for better understanding the associations of bacterioplankton communities in freshwater lakes.

In a comparative study, Steffen *et al.* highlighted the utility of Metagenomics as a tool for exploration of microbial communities, provided microbial snapshots of three separate toxic cyanobacterial blooms, Lake Erie (North America), Lake Tai (Taihu, China), and Grand Lakes, St. Marys (OH, USA), using comparative Metagenomics [21]. They concluded that despite being single samples, these metagenomes provided a unique snapshot of the microbial community associated with toxic cyanobacterial blooms. They noticed that sequences of the Microcystis phage Ma-LMM01 were detected at all three lakes. This was especially worth noting due to the importance of phage in bloom dynamics and termination. Their findings included the presence of the mlrC gene in both Taihu and Erie. This gene is involved in microbial degradation of microcystin, and its presence warrants further inquiry into the presence of potential important microcystin degraders in these lakes. Within their observations key functional genes, such as those involved in nitrogen assimilation, appeared to be more informative than standard 16S rDNA gene analysis and demonstrated that within two similar biological events (blooms in Lake Erie and Taihu) the analogous processes were likely carried out by different members of the community. With this approach, they were able to identify potentially divergent pathways of assimilated nitrogen through the microbial communities of three different blooms. The genomic contribution of heterotrophic bacteria to nitrogen assimilation in Taihu represented a potentially critical contribution of heterotrophic bacteria in driving toxic freshwater blooms.

4.2. Environmental Variables and the Microbial Community Structures

Cao *et al.* conducted a study in 21 freshwater lakes in Yunnan Province, China [14]. In their study, two hypothesized structural equation models were used to explore the bacterial community structure dynamics responding to environmental variables in the investigated plateau lakes. The models highlighted the role of the physical environment, land use, lake morphology and nutrients influencing the bacterial community structure in the ecological processes. Water transparency was demonstrated to be a major driving force in determining the taxon composition of the bacterial community. In contrast with what had been presented in the response of the cyanobacteria community to lake morphology, a relatively weak relationship between the bacterioplankton community and lake morphology was observed, especially lake depth. In addition, the models also showed that TN was more significant than TP for determining the bacterioplankton community structure. The threshold analyses for nonlinear responses suggested substantial changes of the bacterioplankton community structure were strikingly observed at 7.36 for pH and at 25.6% for the percentage of the agricultural area, while the distinct change point of the cyanobacteria community structure responding to pH was at 7.74. Finally, following analyses indicated that there was an apparent shift in dominance from Proteobacteria to Cyanobacteria with increasing nutrient loads. Actinobacteria and Bacteroidetes were induced a sharp decrease and increase crossing the change point along the gradient of the agricultural area.

A study undertook by Touzet *et al.* investigated the dynamics in summer diversity of planktonic cyanobacterial communities and microcystin toxin concentrations in two inter-connected lakes from the west of Ireland, Lough Corrib and Ballyquirke Lough [19]. Phytoplankton biomass was estimated through chlorophyll-a analysis, and Cyanobacteria community fingerprinting was examined by 16S rDNA DGGE analysis. Analyzed quantitative variables included temperature, Secchi depth, chlorophyll-a concentration, dissolved inorganic nitrogen and phosphorus, microcystin concentrations and DGGE-based estimate of cyanobacterial abundance. They observed community change throughout the summer, and identified cyanobacterial genotypes both unique and shared to both lakes. Microcystin concentrations were greater in August than in July and June in both lakes. They indicated that this was concomitant to the increased occurrence of Microcystis as evidenced by DGGE band excision and subsequent sequencing and BLAST analysis. RFLP analysis of PCR amplified mcy-A/E genes clustered together the August samples of both lakes, highlighting a potential change in microcystin producers across the two lakes. The multiple factor analysis of the combined environmental data set for the two lakes highlighted the expected pattern opposing greater water temperature and chlorophyll concentration against macronutrient concentrations, but also indicated a negative relationship between microcystin concentration and cyanobacterial diversity, possibly underlining allelopathic interactions. Despite some element of connectivity, the dissimilarity in the composition of the cyanobacterial assemblages and the timing of community change in the two lakes likely were a reflexion of niche differences determined by meteorologically-forced variation in physico-chemical parameters in the two water bodies.

Using weekly data from western Lake Erie in 2014, Berry *et al.* investigated how the cyanobacterial community varied over space and time, and whether the bloom affected non-cyanobacterial (nc-bacterial) diversity and composition [24]. In the study, extracted DNA was amplified using primer set 515f/806r, which targets the V4 hypervariable regions of the 16S rRNA gene. Both microbial community parameters and environmental parameters were examined in the study. They found that bacterial community exhibited changes in diversity and composition during the bloom season, the evenness of Alphaproteobacteria and Betaproteobacteria showed differential responses to algal pigment levels, suggesting that the bloom affected niche diversity for these phylogenetic groups. Their observations supported a link between CHABs and disturbances to bacterial community diversity and composition. They concluded that changes in community composition could be represented in three coordinates, with the first coordinate associated most strongly with bloom measures, the second coordinate associated with temperature, and the third coordinate associated with physical water mass movements. These results supported work by others demonstrating that bacterial communities are impacted by CHABs, and identifies the acI clade as a particularly affected group. The short recovery of many taxa after the bloom indicates that bacterial communities may exhibit resilience to CHABs.

Tromas *et al.* used a deep 16S amplicon sequencing approach to profile the 32 bacterial communities in eutrophic Lake Champlain over time, to characterize the composition and repeatability of cyanobacterial blooms, and to determine the potential for blooms to be predicted based on time-course sequence data [25]. The analysis, based on 143 samples between 2006 and 2013, spans multiple bloom events. They found that the microbial community varied substantially over months and seasons, while remaining stable from year to year. Bloom events significantly altered the bacterial community but did not reduce overall diversity, suggesting that a distinct microbial community—including noncyano-bacteria—prospers during the bloom. Blooms tended to be dominated by one or two genera of cyanobacteria: *Microcystis* or *Dolichospermum*. Blooms were thus relatively repeatable at the genus level, but more unpredictable at finer taxonomic scales. They classified their samples into bloom or non-bloom bins, achieving up to 92% accuracy. They confirmed that cyanobacterial blooms respond significantly to total phosphorus and total nitrogen as previously described. Temperature was also an important factor shaping the lake microbial community, as previously documented. However, in this study, they observed that these predictors explained only a part of the variation between bloom and no-bloom samples. Other predictors might include water column stability and mixing, and the interactions of predictors, especially nutrients and temperature. In addition to environmental factors, they showed that biological factors, in the form of bacterial OTUs or genera, could also help to characterize the bloom. They indicated that Cyanobacterial blooms alter the local environment, likely altering the surrounding microbial community. As a result, these assemblages likely included bacteria that were reliant on cyanobacterial metabolites and biomass. Using symbolic regression, they were able to predict the start date of a bloom with 78-91% explained variance over tested data (depending on the data used for model training). They stated that sequence data appeared to be a strong predictor, similar or better than prediction with environmental variables. This showed that, although blooms in Lake Champlain (and other temperate lakes) were clearly correlated with seasonality (*i.e.* blooms occur mainly during summer, at warmer temperatures), the state of the microbial community may contain more information than environmental factors alone about the likelihood of an impending bloom. This could be because one microbial taxon contains information about numerous environmental parameters, resulting in parsimonious predictive models based on a small number of taxonomic biomarkers.

4.3. *Microcystis* Appears in Most of the Studies Lakes

Microcystins (MCs) are the most common and potent cyanotoxins in freshwater systems worldwide. The most of the lakes examined in this section 4 were associated with this genus. Lake Taihu has experienced *Microcystis* bloom events for decade, Xie *et al.* [10] and Steffen *et al.* [21] focused their studies on this lake; Cao *et al.* [14] reported that analysis at the genus level of *Cyanobacteria* identified that Microcystis was among the most abundant genus in the 21 plateau lakes in Yunnan, China; Steffen *et al.* [21] stated that *Microcystis*-dominated blooms had been observed in the western basin of Lake Erie annually since the 1990s; the study by Berry *et al.* [24] in western Lake Erie also showed that Cyanobacterial community composition fluctuated dynamically during the bloom, but was dominated by Microcystis and Synechococcus OTUs; Tromas *et al.* [25] reported that blooms in their study site (Lake Champlain, North America) tended to be dominated by one or two genera of cyanobacteria: *Microcystis* or *Dolichospermum*; in the study of Touzet [19] *et al. Microcystins* were extracted from the samples of Lough Corrib and Ballyquirke Lough. Microcystins (MCs) are predominantly produced by *Microcystis* spp. which is considered a serious health hazard due to its potent liver toxicity and carcinogenic potential, and has been seriously concerned.

5. Machine Learning (ML) Approaches Possess Dual Significance in the Metagenomics and Cyanobacteria Blooming Study

Metagenomic data analyses aims at identifying the taxonomic composition of microbes and their relative counts and annotating the functional roles as encoded by micro biomes and finding association of microbes with their functional metadata phenotypes [55] [56]. Differentiate between microbial communities or associated functional conditions can be realized through analysing relative OTU abundance across metagenomic samples and their relationships. Machine learning (ML) techniques are used as a powerful tool in the metagenomic data analyses [57] [58], as it depends on computational tools for analysing sheer data sets, gaining information from the microbial community. This is reflected in the afore-mentioned case studies. Whilst in the early studies focusing on the environmental parameters, machine learning was applied to establish relationships between physical-chemical factors and the blooming occurrence. Therefore, machine learning (ML) has dual implications in the cyanobacteria blooming study. Researchers wish to improve the methodology used both in the metagenomic analysis and in physical-chemical oriented ML modeling approaches as well.

6. Summary Remarks

Cyanobacteria blooms studies have been undertaken for decades, concerning the harmfulness of the blooming to environment and human beings. Conventional studies mainly focus on investigating the environmental factors influencing the blooms, possessing their limitation in lack of viewing the microbial population of the blooming. Metagenomics study provides insight into the internal community structure of the cyanobacteria at the blooming, and there researchers reported that sequence data was a better predictor than environmental factors. This

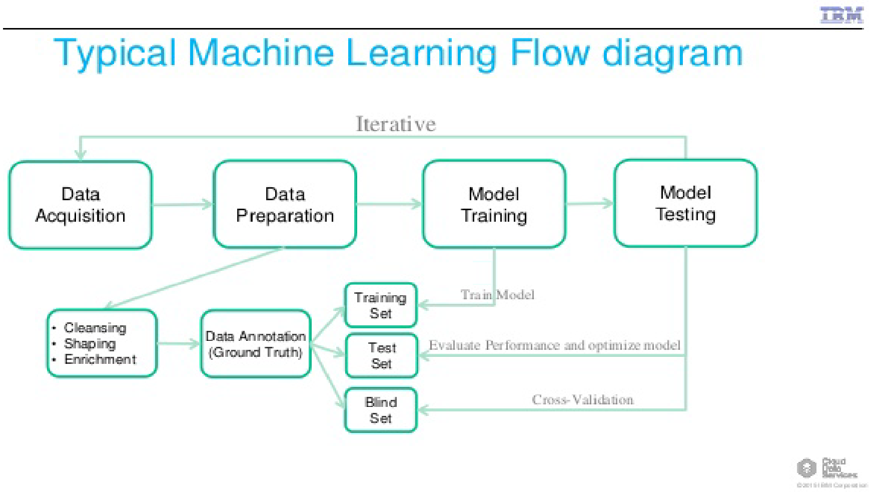

Figure 2. A diagram of machine learning (https://www.google.co.uk/search?q=machine+learning+diagram&tbm=isch&tbo=u&source=univ&sa=X&sqi=2&ved=0ahUKEwiZgsrvu6bUAhXMYVAKHYepAuMQsAQIQA&biw=1366&bih=634).

further manifests the significance of the metagenomic study. However, large number of the latter appears to be confined only to present snapshoot of the microbial community diversity and structure. This type of investigation has been valuable and important, whilst an effort to integrate and coordinate the conventional approaches that largely focus on the environmental factors control, and the metagenomic approaches that reveals the microbial community structure and diversity, implemented through machine learning techniques, for a holistic and more comprehensive insight into the cause and control of Cyanobacteria blooms, appear to be a trend and challenge of the study of this field.

Acknowledgements

This study was part of the "Wuhan lakes" project funded by Ulster University Global Challenges Research Pump-Priming Fund.

References

[1] Zurawell, R.W., Chen, H., Burke, J.M. and Prepas, E.E. (2005) Hepatotoxic Cyanobacteria: A Review of the Biological Importance of Microcystins in Freshwater Environments. *Journal of Toxicology and Environmental Health* (*Part B*), **8**, 1-37. https://doi.org/10.1080/10937400590889412

[2] Carmichael, W.W. and Boyer, G.L. (2016). Health Impacts from Cyanobacteria Harmful Algae Blooms: Implications for the North American Great Lakes. *Harmful Algae*, **54**, 194-212. https://doi.org/10.1016/j.hal.2016.02.002

[3] Bullerjahn, G.S., McKay, R.M., Davis, T.W., Baker, D.B., Boyer, G.L., D'Anglada, L.V., Doucette, G.J., Ho, J.C., Irwin, E.G., Kling, C.L. and Kudela, R.M. (2016) Global Solutions to Regional Problems: Collecting Global Expertise to Address the Problem of Harmful Cyanobacterial Blooms. A Lake Erie Case Study. *Harmful Algae*, **54**, 223-238. https://doi.org/10.1016/j.hal.2016.01.003

[4] Harke, M.J., Steffen, M.M., Gobler, C.J., Otten, T.G., Wilhelm, S.W., Wood, S.A.

and Paerl, H.W. (2016) A Review of the Global Ecology, Genomics, and Biogeography of the Toxic Cyanobacterium, Microcystis spp. *Harmful Algae*, **54**, 4-20. https://doi.org/10.1016/j.hal.2015.12.007

[5] Li, D., Kong, F., Shi, X., Ye, L., Yu, Y. and Yang, Z. (2012) Quantification of Microcystin-Producing and Non-Microcystin Producing Microcystis Populations during the 2009 and 2010 Blooms in Lake Taihu Using Quantitative Real-Time PCR. *Journal of Environmental Sciences*, **24**, 284-290. https://doi.org/10.1016/S1001-0742(11)60745-6

[6] Cai, Y., Kong, F., Shi, L. and Yu, Y. (2012) Spatial Heterogeneity of Cyanobacterial Communities and Genetic Variation of Microcystis Populations within Large, Shallow Eutrophic Lakes (Lake Taihu and Lake Chaohu, China). *Journal of environmental sciences*, **24**, 1832-1842. https://doi.org/10.1016/S1001-0742(11)61007-3

[7] Jia, Y., Dan, J., Zhang, M. and Kong, F. (2013) Growth Characteristics of Algae during Early Stages of Phytoplankton Bloom in Lake Taihu, China. *Journal of Environmental Sciences*, **25**, 254-261. https://doi.org/10.1016/S1001-0742(12)60058-8

[8] Li, W., Qin, B. and Zhu, G. (2014) Forecasting Short-Term Cyanobacterial Blooms in Lake Taihu, China, Using a Coupled Hydrodynamic-Algal Biomass Model. *Ecohydrology*, **7**, 794-802. https://doi.org/10.1002/eco.1402

[9] Huang, X., Hu, B., Wang, P., Chen, X. and Xu, B. (2016) Microbial Diversity in Lake-River Ecotone of Poyang Lake, China. *Environmental Earth Sciences*, **75**, 1-7. https://doi.org/10.1007/s12665-016-5473-0

[10] Xie, M., Ren, M., Yang, C., Yi, H., Li, Z., Li, T. and Zhao, J. (2016) Metagenomic Analysis Reveals Symbiotic Relationship among Bacteria in Microcystis-Dominated Community. Frontiers in Microbiology, **7**. https://doi.org/10.3389/fmicb.2016.00056

[11] Yabunaka, K.I., Hosomi, M. and Murakami, A. (1997) Novel Application of a Back-Propagation Artificial Neural Network Model Formulated to Predict Algal Bloom. *Water Science and Technology*, **36**, 89-97. https://doi.org/10.1016/S0273-1223(97)00464-2

[12] Zhou, J., Qin, B., Han, X. and Zhu, L. (2016) Turbulence Increases the Risk of Microcystin Exposure in a Eutrophic Lake (Lake Taihu) during Cyanobacterial Bloom periods. *Harmful Algae*, **55**, 213-220. https://doi.org/10.1016/j.hal.2016.03.016

[13] Zhang, J., Zhu, C., Guan, R., Xiong, Z., Zhang, W., Shi, J., Sheng, Y., Zhu, B., Tu, J., Ge, Q. and Chen, T. (2017) Microbial Profiles of a Drinking Water Resource based on Different 16S rRNA V regions during a Heavy Cyanobacterial Bloom in Lake Taihu, China. *Environmental Science and Pollution Research*, **24**, 12796-12808. https://doi.org/10.1007/s11356-017-8693-2

[14] Cao, X., Wang, J., Liao, J., Gao, Z., Jiang, D., Sun, J., Zhao, L., Huang, Y. and Luan, S. (2017) Bacterioplankton Community Responses to Key Environmental Variables in Plateau Freshwater Lake Ecosystems: A Structural Equation Modeling and Change Point Analysis. *Science of the Total Environment*, **580**, 457-467. https://doi.org/10.1016/j.scitotenv.2016.11.143

[15] Eiler, A. and Bertilsson, S. (2004) Composition of Freshwater Bacterial Communities Associated with Cyanobacterial Blooms in Four Swedish Lakes. *Environmental Microbiology*, **6**, 1228-1243. https://doi.org/10.1111/j.1462-2920.2004.00657.x

[16] Louati, I., Pascault, N., Debroas, D., Bernard, C., Humbert, J.F. and Leloup, J. (2015) Structural Diversity of Bacterial Communities Associated with Bloom-Forming Freshwater Cyanobacteria Differs According to the Cyanobacterial Genus. *PloS one*, **10**. https://doi.org/10.1371/journal.pone.0140614

[17] Louati, I., Pascault, N., Debroas, D., Bernard, C., Humbert, J.F. and Leloup, J. (2016)

Correction: Structural Diversity of Bacterial Communities Associated with Bloom-Forming Freshwater Cyanobacteria Differs According to the Cyanobacterial Genus. *PloS one*, **11**. https://doi.org/10.1371/journal.pone.0146866

[18] Winter, C., Hein, T., Kavka, G., Mach, R.L. and Farnleitner, A.H. (2007) Longitudinal Changes in the Bacterial Community Composition of the Danube River: A Whole-River Approach. *Applied and Environmental Microbiology*, **73**, 421-431. https://doi.org/10.1128/AEM.01849-06

[19] Touzet, N., McCarthy, D., Gill, A. and Fleming, G.T.A. (2016) Comparative Summer Dynamics of Surface Cyanobacterial Communities in Two Connected Lakes from the West of Ireland. *Science of the Total Environment*, **553**, 416-428. https://doi.org/10.1016/j.scitotenv.2016.02.117

[20] Graham, J.L., Jones, J.R., Jones, S.B., Downing, J.A. and Clevenger, T.E. (2004) Environmental Factors Influencing Microcystin Distribution and Concentration in the Midwestern United States. *Water research*, **38**, 4395-4404. https://doi.org/10.1016/j.watres.2004.08.004

[21] Steffen, M.M., Li, Z., Effler, T.C., Hauser, L.J., Boyer, G.L. and Wilhelm, S.W. (2012) Comparative Metagenomics of Toxic Freshwater Cyanobacteria Bloom Communities on Two Continents. *PloS one*, **7**. https://doi.org/10.1371/journal.pone.0044002

[22] Gobler, C.J., Burkholder, J.M., Davis, T.W., Harke, M.J., Johengen, T., Stow, C.A. and Van de Waal, D.B. (2016) The Dual Role of Nitrogen Supply in Controlling the Growth and Toxicity of Cyanobacterial Blooms. *Harmful Algae*, **54**, 87-97. https://doi.org/10.1016/j.hal.2016.01.010

[23] Watson, S.B., Miller, C., Arhonditsis, G., Boyer, G.L., Carmichael, W., Charlton, M.N., Confesor, R., Depew, D.C., Höök, T.O., Ludsin, S.A. and Matisoff, G. (2016) The Re-Eutrophication of Lake Erie: Harmful Algal Blooms and Hypoxia. *Harmful Algae*, **56**, 44-66. https://doi.org/10.1016/j.hal.2016.04.010

[24] Berry, M.A., Davis, T.W., Cory, R.M., Duhaime, M.B., Johengen, T.H., Kling, G.W., Marino, J.A., Den Uyl, P.A., Gossiaux, D., Dick, G.J. and Denef, V.J. (2017) Cyanobacterial Harmful Algal Blooms are a Biological Disturbance to Western Lake Erie Bacterial Communities. *Environmental Microbiology*, **19**, 1149-1162. https://doi.org/10.1111/1462-2920.13640

[25] Tromas, N., Fortin, N., Bedrani, L., Terrat, Y., Cardoso, P., Bird, D., *et al.* (2016) Characterizing and Predicting Cyanobacterial Blooms in an 8-Year Amplicon Sequencing Time-Course. *bioRxiv*.

[26] Ho, J.C. and Michalak, A.M. (2015) Challenges in Tracking Harmful Algal Blooms: A Synthesis of Evidence from Lake Erie. *Journal of Great Lakes Research*, **41**, 317-325. https://doi.org/10.1016/j.jglr.2015.01.001

[27] Pope, P.B. and Patel, B.K. (2008) Metagenomic Analysis of a Freshwater Toxic Cyanobacteria Bloom. *FEMS Microbiology Ecology*, **64**, 9-27. https://doi.org/10.1111/j.1574-6941.2008.00448.x

[28] Woodhouse, J.N., Kinsela, A.S., Collins, R.N., Bowling, L.C., Honeyman, G.L., Holliday, J.K. and Neilan, B.A. (2015) Microbial Communities Reflect Temporal Changes in Cyanobacterial Composition in a Shallow Ephemeral Freshwater Lake. *The ISME Journal*.

[29] Sitoki, L., Kurmayer, R. and Rott, E. (2012) Spatial Variation of Phytoplankton Composition, Biovolume, and Resulting Microcystin Concentrations in the Nyanza Gulf (Lake Victoria, Kenya). *Hydrobiologia*, **691**, 109-122. https://doi.org/10.1007/s10750-012-1062-8

[30] Ndlela, L.L., Oberholster, P.J., Van Wyk, J.H. and Cheng, P.H. (2016) An Overview of Cyanobacterial Bloom Occurrences and Research in Africa over the Last Decade.

Harmful Algae, **60**, 11-26. https://doi.org/10.1016/j.hal.2016.10.001

[31] Wu, S.K., Xie, P., Liang, G.D., Wang, S.B. and Liang, X.M. (2006) Relationships between Microcystins and Environmental Parameters in 30 Subtropical Shallow Lakes along the Yangtze River, China. *Freshwater Biology*, **51**, 2309-2319. https://doi.org/10.1111/j.1365-2427.2006.01652.x

[32] Conley, D.J., Paerl, H.W., Howarth, R.W., Boesch, D.F., Seitzinger, S.P., Havens, K.E., Lancelot, C. and Likens, G.E. (2009) Controlling Eutrophication: Nitrogen and Phosphorus. *Science*, **323**, 1014-1015. https://doi.org/10.1126/science.1167755

[33] Davis, T.W., Berry, D.L., Boyer, G.L. and Gobler, C.J. (2009) The Effects of Temperature and Nutrients on the Growth and Dynamics of Toxic and Non-Toxic Strains of Microcystis during Cyanobacteria Blooms. *Harmful algae*, **8**, 715-725. https://doi.org/10.1016/j.hal.2009.02.004

[34] Pace, N.R., Stahl, D.A., Lane, D.J. and Olsen, G.J. (1986) The Analysis of Natural Microbial Populations by Ribosomal RNA Sequences. *Advances in Microbial Ecology*, **9**, 1-55. https://doi.org/10.1007/978-1-4757-0611-6_1

[35] Winter, C., Hein, T., Kavka, G., Mach, R.L. and Farnleitner, A.H. (2007) Longitudinal Changes in the Bacterial Community Composition of the Danube River: A Whole-River Approach. *Applied and Environmental Microbiology*, **73**, 421-431. https://doi.org/10.1128/AEM.01849-06

[36] Tan, B., Ng, C.M., Nshimyimana, J.P., Loh, L.L., Gin, K.Y.H. and Thompson, J.R. (2015) Next-Generation Sequencing (NGS) for Assessment of Microbial Water Quality: Current Progress, Challenges, and Future Opportunities. *Frontiers in Microbiology*, **6**, 1027. https://doi.org/10.3389/fmicb.2015.01027

[37] Le, C., Zha, Y., Li, Y., Sun, D., Lu, H. and Yin, B. (2010) Eutrophication of Lake Waters in China: Cost, Causes, and Control. *Environmental Management*, **45**, 662-668. https://doi.org/10.1007/s00267-010-9440-3

[38] Qin, B., Zhu, G., Gao, G., Zhang, Y., Li, W., Paerl, H.W. and Carmichael, W.W. (2010) A Drinking Water Crisis in Lake Taihu, China: Linkage to Climatic Variability and Lake Management. *Environmental Management*, **45**, 105-112. https://doi.org/10.1007/s00267-009-9393-6

[39] Kong, F. and Fao, G. (2005) Hypothesis on Cyanobacteria Bloom-Forming Mechanism in Large Shallow Eutrophic Lakes. *Acta ecologica sinica/Shengtai Xuebao*, **25**, 589-595.

[40] Wilhelm, S.W., Farnsley, S.E., LeCleir, G.R., Layton, A.C., Satchwell, M.F., DeBruyn, J.M., Boyer, G.L., Zhu, G. and Paerl, H.W. (2011) The Relationships between Nutrients, Cyanobacterial Toxins and the Microbial Community in Taihu (Lake Tai), China. *Harmful Algae*, **10**, 207-215. https://doi.org/10.1016/j.hal.2010.10.001

[41] McCarthy, M.J., Lavrentyev, P.J., Yang, L., Zhang, L., Chen, Y., Qin, B. and Gardner, W.S. (2007) Nitrogen Dynamics and Microbial Food Web Structure during a Summer Cyanobacterial Bloom in a Subtropical, Shallow, Well-Mixed, Eutrophic lake (Lake Taihu, China). *Hydrobiologia*, **581**, 195-207. https://doi.org/10.1007/s10750-006-0496-2

[42] Ma, J., Brookes, J.D., Qin, B., Paerl, H.W., Gao, G., Wu, P., Zhang, W., Deng, J., Zhu, G., Zhang, Y. and Xu, H. (2014) Environmental Factors Controlling Colony Formation in Blooms of the Cyanobacteria Microcystis spp. in Lake Taihu, China. *Harmful algae*, **31**, 136-142. https://doi.org/10.1016/j.hal.2013.10.016

[43] Krausfeldt, L.E., Tang, X., van de Kamp, J., Gao, G., Bodrossy, L., Boyer, G.L. and Wilhelm, S.W. (2017) Spatial and Temporal Variability in the Nitrogen Cyclers of Hypereutrophic Lake Taihu. *FEMS Microbiology Ecology*, **93**. https://doi.org/10.1093/femsec/fix024

[44] Muttil, N. and Chau, K.W. (2006) Neural Network and Genetic Programming for Modelling Coastal Algal Blooms. *International Journal of Environment and Pollution*, **28**, 223-238. https://doi.org/10.1504/IJEP.2006.011208

[45] Khataee, A.R., Dehghan, G., Zarei, M., Ebadi, E. and Pourhassan, M. (2011) Neural Network Modeling of Biotreatment of Triphenylmethane Dye Solution by a Green Macroalgae. *Chemical Engineering Research and Design*, **89**, 172-178. https://doi.org/10.1016/j.cherd.2010.05.009

[46] Vilán, J.V., Fernández, J.A., Nieto, P.G., Lasheras, F.S., de Cos Juez, F.J. and Mu-iz, C.D. (2013) Support Vector Machines and Multilayer Perceptron Networks Used to Evaluate the Cyanotoxins Presence from Experimental Cyanobacteria Concentrations in the Trasona Reservoir (Northern Spain). *Water Resources Management*, **27**, 3457-3476. https://doi.org/10.1007/s11269-013-0358-4

[47] Handelsman, J., Rondon, M.R., Brady, S.F., Clardy, J. and Goodman, R.M. (1998) Molecular Biological Access to the Chemistry of Unknown Soil Microbes: A New Frontier for Natural Products. *Chemistry & Biology*, **5**, R245-R249. https://doi.org/10.1016/s1074-5521(98)90108-9

[48] Pinto, A.J., Xi, C. and Raskin, L. (2012) Bacterial Community Structure in the Drinking Water Microbiome Is Governed by Filtration Processes. *Environmental Science & Technology*, **46**, 8851-8859. https://doi.org/10.1021/es302042t

[49] Sanger, F., Nicklen, S. and Coulson, A.R. (1977) DNA Sequencing with Chain-Terminating Inhibitors. *Proceedings of the National Academy of Sciences*, 5463-5467. https://doi.org/10.1073/pnas.74.12.5463

[50] Schuster, S.C. (2008) Next-Generation Sequencing Transforms Today's Biology. *Nature Methods*, **5**, 16. https://doi.org/10.1038/nmeth1156

[51] Mardis, E.R. (2008) The Impact of Next-Generation Sequencing Technology on Genetics. *Trends in Genetics*, **24**, 133-141. https://doi.org/10.1016/j.tig.2007.12.007

[52] Ansorge, W.J. (2009) Next-Generation DNA Sequencing Techniques. *New Biotechnol*, **25**, 195-203. https://doi.org/10.1016/j.nbt.2008.12.009

[53] Hugenholtz, P. and Tyson, G.W. (2008) Microbiology: Metagenomics. *Nature*, **455**, 481-483. https://doi.org/10.1038/455481a

[54] Zhao, D., Shen, F., Zeng, J., Huang, R., Yu, Z. and Wu, Q.L. (2016) Network Analysis Reveals Seasonal Variation of Co-occurrence Correlations between Cyanobacteria and Other Bacterioplankton. *Science of the Total Environment*, **573**, 817-825. https://doi.org/10.1016/j.scitotenv.2016.08.150

[55] Marco, D., Ed. (2010) Metagenomics: Theory, Methods and Applications. Horizon Scientific Press.

[56] Wooley, J.C., Godzik, A. and Friedberg, I. (2010) A Primer on Metagenomics. *PLoS Comput Biol*, **6**. https://doi.org/10.1371/journal.pcbi.1000667

[57] Bouchot, J.L., Trimble, W.L., Ditzler, G., Lan, Y., Essinger, S. and Rosen, G. (2013) Advances in Machine Learning for Processing and Comparison of Metagenomic Data.

[58] Soueidan, H. and Nikolski, M. (2015) Machine Learning for Metagenomics: Methods and Tools.