**Secondary Metabolites Are Highly Predictive of Diazotrophic Cyanobacteria Strains**

James Young, Liping Gu, Ruanbao Zhou,

South Dakota State University

**Abstract**

Nitrogen fixation (diazotrophy) is a desirable trait present in some cyanobacteria that has garnered attention for its potential applications in sustainable agriculture and chemical feedstock production (Young, 2021 #457). Different strains of cyanobacteria have different diazotrophic strategies and finding new diazotrophic strains may result in improved industrial application and fundamental discovery (Young, 2021 #457). In this paper we use chemical structure similarity of secondary metabolites to predict whether a cyanobacteria is diazotrophic or non-diazotrophic. The model achieved 89% accuracy and ROC-AUC of 0.97 on holdout data through leave one out cross validation on 158 manually labeled metabolites. This modeling approach was used to predict whether the remaining ~2000 chemicals were likely to be associated with diazotrophic cyanobacteria or not, culminating in a list of strains suspected to be diazotrophic. Broader questions are considered regarding why secondary metabolites can be highly predictive of diazotrophic strains. By taking the approach of using secondary metabolites rather than genetic information to predict diazotrophic cyanobacteria, this work creates a few situational advantages. One advantage is the possibility of discovering diazotrophs even if they are not actively fixing nitrogen during environmental sampling. Another advantage is prioritizing work on strains where there is limited genetic sequencing information available.

**Introduction**

Cyanobacteria are ancient prokaryotes that have evolved to thrive in diverse environments under pressures ranging from extreme heat and cold to bio-available nitrogen deprivation (Bahl, 2011 #451). In addition to having unique metabolic pathways of photosynthesis and sometimes nitrogen fixation, cyanobacteria are also a rich source of secondary metabolites (Nunnery, 2010 #452). Secondary metabolites are chemicals produced by the organism that are not directly necessary for building new cells or creating energy, but rather carry out roles such as anti-bacterial effects or protease inhibition, which can help the cell (Singh, 2011 #458). Secondary metabolites are also known for their medicinal effects, making them targets of interest for biotechnological development (Nunnery, 2010 #452).

Prior work has shown that the habitat of the cyanobacteria is strongly associated with different secondary metabolites (Monteiro, 2021 #453). This work extends this finding in a mathematically rigorous way by finding associations between secondary metabolite chemical structure and the likelihood that it was produced by a diazotrophic cyanobacteria. We make use of atom-pair fingerprint similarity of chemical structure through chemmineR (Cao, 2008 #454) to build a probability that the chemical came from a diazotrophic cyanobacteria. This approach is similar to the well-developed field of quantitative structure activity relationship (Dudek, 2006 #455), with the twist that we are predicting something about the cell producing the chemical, rather than the activity of the chemical itself.

**Methods**

Data

Data was taken from CyanoMetDB (Jones, 2021 #456). Strains of cyanobacteria within the database were manually checked on UniProt for having a NifH/NifD/NifK protein associated with it as a marker for nitrogen fixation. Strains having one or more of these proteins present in their proteome were marked “1” for nitrogen-fixing, and strains without any of these proteins were marked “0” for not nitrogen fixing. If a strain did not have a sufficient proteome size (> 1000) then it was not marked 0 or 1 and left as unlabeled.

CyanoMetDB was filtered to have only chemicals associated with labeled strains to train and evaluate the model. The model was built in R by building an sdfset of the SMILES chemical codes, the sdfset was then parsed into atom pair counts which numerically represent its associated chemical structure. The model then finds the most similar atom-pair representation to the chemical you are querying, gives it’s similarity score (between 0 and 1) and the class of that database chemical.

The formula of modeling the probability was conditional on whether the most similar labeled metabolite was associated with a diazotrophic or non-diazotrophic cyanobacteria.

If the most similar metabolite was diazotroph associated, then the formula is:

Probability(Diazotrophic =True ) = 0.5 + (0.5 x similarity to most similar chemical)

If the most similar metabolite was non-diazotroph associated, then the formula is:

Probability(Diazotrophic =True ) = 0.5 - (0.5 x similarity to most similar chemical)

Where the maximum similarity is 1 and the minimum was 0.01.

The model evaluation was done using leave one out cross validation (LOOCV) for both the accuracy evaluation and the ROC-AUC evaluation.

**Results**

**Predictive Power**

Our atomic-pair based structural similarity model predicted the correct class (diazotrophic or non-diazotrophic) with 89.9% accuracy and a ROC-AUC of 0.97 on leave one out cross validation holdouts (158 manually labeled compounds total). If a person were to make the naïve prediction that all chemicals represented the majority class (non-diazotrophic) in the training set, they would only be correct 53.8% of the time. Our lift and response plot shows that when we rank our predictions by probability of being diazotrophic, our top predictions (closer to 1 on the x-axis) improve the odds by approximately 2-fold, which is demonstrated in the lift line. We can also see that our top 60% of predictions capture all the diazotrophic associated metabolites, demonstrated in the cumulative response line. This can be interpreted to mean we can have more confidence in a prediction being correct when it has higher estimated probability. We can use this confidence in higher ranked predictions to prioritize strains to explore and metabolites to trust as indicators of diazotrophs.

Chart, line chart

Description automatically generatedChart

Description automatically generated 

Figure 1: Top left: Accuracy and other metrics demonstrate a substantial improvement over the baseline of 53.8% accuracy. Top right: The predictions made with highest probability are highly accurate and demonstrates usefulness for ranking predictions from most to least likely. Bottom left: The ROC-AUC of the holdouts is near a perfect score with a value of 0.97 further supporting the power of this approaches rank-ordering. Bottom right: 3 chemicals associated with diazotrophic strains: N-(2-aminoethyl)-glycine, Kemibelactone A, and Hoiamide.

**Top Unknown Strains and Metabolites for Exploration**

Applying our method to the unlabeled chemical compounds, we created a ranking of most to least likely secondary metabolites being associated with a diazotrophic cyanobacteria. The top 10 strains are presented in table 1 below. The top ranked strain based on probability of being diazotrophic is IL-208-2-2 from the genus Schizothrix which was isolated in soil. Schizothrix in water is known to be diazotrophic 1 and further supporting this top ranked strain, the top ranked metabolite in table 2 was isolated from this strain. While Schizothrix genus has nifH sequences in UniProt, the specific strain IL-208-2-2 has no sequences in UniProt but we have strong evidence to believe it is diazotrophic based on our model and the supporting evidence.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Strain** | **Max** | **Min** | **Mean** | **Median** | **Count** |
| IL-208-2-2 | 0.961 | 0.961 | 0.961 | 0.961 | 1 |
| TAU NZ-3-1 | 0.955 | 0.891 | 0.925 | 0.928 | 3 |
| AV1 | 0.955 | 0.808 | 0.869 | 0.863 | 17 |
| UIC 10045 | 0.943 | 0.208 | 0.576 | 0.576 | 2 |
| PCC7310 | 0.929 | 0.929 | 0.929 | 0.929 | 1 |
| CENA352 | 0.921 | 0.919 | 0.920 | 0.920 | 2 |
| KAC 11 | 0.919 | 0.795 | 0.840 | 0.823 | 4 |
| BY1 | 0.917 | 0.909 | 0.913 | 0.913 | 2 |

Table 1: The strains shown in the table have at least one metabolite with top 10 similarity to known diazotrophs. From the evaluation on strains of known diazotrophic genetics, we would expect these top ranked strains to be diazotrophs with a high likelihood. The max, min, mean, and median columns refer to those summary statistics of all chemicals associated with the strain. The count is how many chemicals were present for a given strain. Full results are available in the supplemental information.

|  |  |
| --- | --- |
| **Compound Name** | **Probability of Diazotroph** |
| Schizopeptin 791 | 0.961 |
| Anabaenopeptin NZ857 | 0.955 |
| Nodulapeptin B | 0.955 |
| Nodulapeptin 855b | 0.952 |
| Laxaphycin B | 0.950 |
| Trichormamide C | 0.943 |
| Laxaphycin B2 | 0.943 |
| Nodulapeptin 915a | 0.941 |
| Nodulapeptin 863 | 0.934 |
| Laxaphycin B3 | 0.932 |

Table 2: The metabolites shown in the table come from the unlabeled data and are in the top 10 similarity to known diazotroph associated metabolites. Full results are available in the supplemental information.

**Toxicity is Not Positively Associated with Diazotrophic Secondary Metabolites**

The ability of secondary metabolites to accurately predict diazotrophic strains raises the question if they play a general role across diazotrophic species. One known role of cyanobacterial secondary metabolites is their cytotoxicity towards other organisms. Could there be a difference between the toxicity of diazotrophic and non-diazotrophic associated secondary metabolites? Perhaps the diazotrophic cells produce more toxic metabolites to fend of predators from consuming them for nitrogen. We ran the labeled metabolites through EPA’s Toxicity Estimation Software Tool (results in Supplementary Info) for rats and Daphnia magna. Diazotrophic associated metabolites were not significantly more toxic than non-diazotrophic associated metabolites. In rats, contrary to the hypothesis posited above, the diazotrophic associated metabolites were significantly less toxic than non-diazotrophic metabolites when evaluated by a two-sided t test. While not informative in the function of these diazotrophic associated metabolites, this is positive for the use of diazotrophic strains in industrial and agricultural applications.

![Chart, scatter chart

Description automatically generated]()

Figure 2: Above left: Diazotrophic associated metabolites have lower LD50 in rats than non-diazotrophic associated metabolites. Above right: Diazotrophic associated metabolites are not significantly different in LD50 in Daphnia magna than non-diazotrophic associated metabolites.

**Discussion**

Interest in cyanobacterial secondary metabolites for use in biomedical cases has been established for decades (Kini, 2020 #459). However, the systematic and statistical use of cyanobacterial metabolites as indicators for cyanobacteria with nitrogen-fixing enzymes has not been considered prior to this work. We find that chemical structure similarity, as determined by atom-pair fingerprints, is a strong predictor variable for the likelihood of the producing organisms being diazotrophic. The usefulness of this work is its allowance for predicting and ranking cyanobacterial strains that don’t have complete proteomes published as suspected diazotrophs based on their metabolites. This can allow for prioritization of strains to sequence and characterize for diazotrophic strength and nitrogenase protection mechanisms.

In addition to the practical application of this work, it has also raised questions regarding why diazotrophic cyanobacteria have structurally similar secondary metabolites. Many cyanobacterial secondary metabolites serve to protect the cells from biological threats through various cytotoxic effects. When we explored the cytotoxicity of secondary metabolites from our training data, no positive association was seen between predicted toxicity level and diazotrophic strains. This is positive news for the use of diazotrophic cyanobacteria in industrial and agricultural processes, but leaves the question open regarding the purpose of these secondary metabolites.

As more metabolites of cyanobacteria are discovered, the understanding of both the level of conservation and purpose these chemicals serve within diazotrophic cyanobacteria should increase. Our model has shown that secondary metabolites are strongly predictive of diazotrophic genetic complements in new cyanobacteria and has opened the question as to why. This also raises the question if these metabolites are present regardless of if the diazotrophic cyanobacteria is currently expressing that phenotype, which will require further refinement of the available databases. If cyanobacteria express their secondary metabolites regardless of current nitrogen status, this approach could be used to identify diazotrophic strains in the environment even when they are not actively fixing nitrogen.

**Supplementary Information**

All supplementary information, including code and data, can be found at <https://github.com/jamesyoung93/SecondaryMetaboliteDiazotrophs>.

**References**

Fix references.

1 Berrendero, E. *et al.* Nitrogen fixation in a non-heterocystous cyanobacterial mat from a mountain river. *Scientific Reports* **6**, 30920, doi:10.1038/srep30920 (2016).