

4-11-2022

Diel gene expression improves software prediction of cyanobacterial operons

Philip Heller

San Jose State University, philip.heller@sjsu.edu

Follow this and additional works at: https://scholarworks.sjsu.edu/faculty_rsca

Recommended Citation

Philip Heller. "Diel gene expression improves software prediction of cyanobacterial operons" *PeerJ* (2022). <https://doi.org/10.7717/peerj.13259>

This Article is brought to you for free and open access by SJSU ScholarWorks. It has been accepted for inclusion in Faculty Research, Scholarly, and Creative Activity by an authorized administrator of SJSU ScholarWorks. For more information, please contact scholarworks@sjsu.edu.

Diel gene expression improves software prediction of cyanobacterial operons

Philip Heller

Department of Computer Science, San Jose State University, San Jose, CA, United States

ABSTRACT

Cyanobacteria are important participants in global biogeochemical process, but their metabolic processes and genomic functions are incompletely understood. In particular, operon structure, which can provide valuable metabolic and genomic insight, is difficult to determine experimentally, and algorithmic operon predictions probably underestimate actual operon extent. A software method is presented for enhancing current operon predictions by incorporating information from whole-genome time-series expression studies, using a Machine Learning classifier. Results are presented for the marine cyanobacterium *Crocospaera watsonii*. A total of 15 operon enhancements are proposed. The source code is publicly available.

Subjects Bioinformatics, Computational Biology, Data Mining and Machine Learning

Keywords Cyanobacteria, Diel, Operon, Crocospaera

INTRODUCTION

Photosynthesizing bacteria (phylum *Cyanobacteria*) are significant participants in global biogeochemical cycles. They arose on earth 3.5 billion years ago (*Schopf, 2002*), and had oxygenated the atmosphere by 2.5 billion years ago (*Tomitani et al., 2006*). Cyanobacteria participate in the ocean biological carbon pump (*Tréguer et al., 2018*), which transports atmospheric greenhouse carbon dioxide to sequestration in the deep ocean. Nitrogen reducing cyanobacteria (diazotrophs) annually convert approximately 200 Tg of atmospheric dinitrogen to bioavailable form (*Wang et al., 2019; Tang, Li & Cassar, 2019*). Cyanobacteria are used to produce medicines (*Tan, 2007; Soni, Trivedi & Madamwar, 2008; Zanchett & Oliveira-Filho, 2013*), biofuels (*Sakurai et al., 2015; Farrokh et al., 2019*), fertilizers (*Singh, Khattar & Ahluwalia, 2014; Chittapun et al., 2018*), cosmetics (*Morone et al., 2019*), and food (*Khan, Bhadouria & Bisen, 2005*).

Despite their ecological and commercial importance, the metabolic processes of many cyanobacteria have not been fully characterized; this is especially true for marine cyanobacteria, which are difficult to cultivate (*Zehr, 2011*). In particular, identification of operons (consecutive genes on the same DNA strand, controlled by a single promoter and expressed as a single transcript) appears to be incomplete. Operon identification provides clues for the inference of regulatory pathways (*Zheng et al., 2002; Westover et al., 2005*), supports interpretation of transcriptome experiments (*Moreno-Hagelsieb & Collado-Vides, 2002*), and can guide annotation of hypothetical genes. The expense of wet lab operon discovery has prompted the development of algorithms for predicting operons from assembled genomes (*Zheng et al., 2002; Moreno-Hagelsieb & Collado-Vides, 2002;*

Submitted 10 September 2021

Accepted 22 March 2022

Published 11 April 2022

Corresponding author

Philip Heller, philip.heller@sjsu.edu

Academic editor

James Procter

Additional Information and
Declarations can be found on
page 6

DOI 10.7717/peerj.13259

© Copyright
2022 Heller

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

Price, 2005); predictions from one of these algorithms (*Price, 2005*) for 1,336 organisms are publicly available (<http://www.microbesonline.org/operons/OperonList.html>). However, few of these predictions have been experimentally verified and it is possible that operon sizes have been underestimated.

Information for honing *in silico* operon predictions can be extracted from time-series measurements of gene expression. Many cyanobacterial genes are not expressed at constant rates, but rather exhibit fluctuating transcript abundance in repeating patterns over a 24-h cycle. For example, production of light-harvesting photosystem II proteins, which are only useful during daylight and whose half-lives are generally less than 12 h (*Yao et al., 2012; Renger et al., 1989*), approximately coincides with available light (*Dodd, 2005*). Since oxygen disables nitrogenase (the enzyme responsible for nitrogen fixation), diazotrophic cyanobacteria segregate nitrogenase from the oxygen evolved by photosynthesis (*Bergersen, 1962; Fay, 1992*); segregation is sometimes temporal, with nitrogenase component proteins produced hours out of phase from photosystem II proteins (*Tuit, Waterbury & Ravizza, 2004*). Diel cycling, defined as a transcript abundance change of at least 2× over 24 h, has been observed in 79% of genes of the diazotrophic cyanobacterium *Crocospaera watsonii* (*Shi et al., 2010*). Since genes in an operon are expected to have similar expression signatures (*Sabatti, 2002; Lercher, 2003*), a high degree of diel expression similarity among adjacent genes might indicate operon membership. Thus if two predicted operons are adjacent, are on the same DNA strand, and exhibit similar diel expression, then the predicted operons may in fact belong to a single common operon.

The approach presented here uses a machine learning classifier – specifically a logistic model tree (*Landwehr, Hall & Frank, 2005; Sumner, Frank & Hall, 2005*) (LMT) – to determine when predicted operons in *Crocospaera* should be merged. A common metric for quantifying expression similarity is Pearson’s Correlation Coefficient (PCC); however, our earlier work (*Heller & Baiju, 2018*) has determined that PCC has deficiencies when applied to the current problem. The “Area Between Linear Interpolations of Measurements” (ABLIM) metric, which we have presented elsewhere (*Heller & Baiju, 2018*), is more appropriate and is the basis of the research reported here. Based on the ABLIM metric, positive and negative example operons were located in the *Crocospaera watsonii* genome. A total of 45 kinds of classifier ([Table S1](#)) were evaluated, and LMT was selected due to its high accuracy. Adjacent predicted operons were identified as candidates for merging, and the expression similarity of all genes was analyzed by the classifier. A total of 15 pairs of candidate operon predictions are recommended for merging ([Table 1](#)).

Source code and instructions are available at <https://github.com/PhilipHeller/Operons> (DOI 10.5281/zenodo.5759925).

MATERIALS AND METHODS

Computed operon predictions (hereafter the “prior predictions”) for strain *Crocospaera watsonii* were downloaded from <http://www.microbesonline.org/operons/>. Log-expression measurements for 4,407 *Crocospaera* genes with eight timepoints were retrieved from a study by *Shi et al. (2010)*. For each gene, log-expressions were normalized to a mean of

Table 1 Pairs of predicted operons recommended for merging.

| Score | Prior 1 | Prior 1 gene functions | Prior 2 | Prior 2 gene functions |
|-------|-------------|--|-------------|---|
| 0.868 | 2207 | HAD-superfamily hydrolase, | 2209 | No predicted function |
| | 2208 | subfamily IA, var No predicted function | 2210 | No predicted function |
| 0.834 | 3078 | No predicted function | 3080 | Glucose-6-phosphate dehydrogenase |
| | 3079 | No predicted function | 3081 | OpcA |
| 0.673 | 4526 | Hydrogenase expression/ synthesis, HypA Hydrogenase | 4528 | No predicted function |
| | 4527 | accessory protein HypB | 4529 | No predicted function |
| 0.672 | 3989 | extracellular solute-binding | 3991 | No predicted function |
| | 3990 | protein, family 3 Amino acid ABC transporter, permease protein | 3992 | No predicted function |
| 0.631 | 5385 | No predicted function | 5388 | Cytochrome-c oxidase |
| | 5386 | Cytochrome c oxidase, subunit II: Cytochrome | 5389 | Cytochrome c oxidase, subunit 3 |
| 0.618 | 1168 | 3-isopropylmalate dehydratase | 1165 | No predicted function |
| | 1169 | small subunit DegT/DnrJ/EryC1/StrS aminotransferase | 1166 | Glutathione S-transferase, N-term |
| 0.599 | 4216 | Pentapeptide repeat | 4214 | No predicted function |
| | 4217 | No predicted function | 4215 | HEAT:PBS lyase HEAT-like rpt |
| 0.576 | 2640 | Phosphopantethiene-protein | 2638 | K ⁺ channel, pore region |
| | 2641 | transferase ATP-binding region, ATPase- like:Histidin | 2639 | K ⁺ channel, pore region |
| 0.565 | 6744 | No predicted function | 6742 | Competence-damaged protein:CinA, C-trmnl |
| | 6745 | No predicted function | 6743 | Uracil phosphoribosyltransferase |
| 0.56 | 4082 | Carbamoyltransferase | 4084 | No predicted function |
| | 4083 | No predicted function | 4085 | No predicted function |
| 0.552 | 2855 | Glycosyl transferase, group 1 | 2853 | Phycobilisome linker polypeptide |
| | 2856 | No predicted function | 2854 | Ferredoxin-dependent bilin reductase |
| 0.537 | 5116 | GTP cyclohydrolase I | 5114 | Cobalamin synthesis protein/P47K:Cobalami |
| | 5117 | Cobalamin synthesis protein/ P47K:Cobalami | 5115 | Dihydrouridine synthase TIM-barrel protein yjbN |
| 0.537 | 2158 | Ribosomal protein L33 | 2160 | Exoribonuclease II |
| | 2159 | Ribosomal protein S18 | 2161 | No predicted function |
| 0.518 | 3464 | Porphobilinogen synthase | 3466 | No predicted function |
| | 3465 | No predicted function | 3467 | Metallophosphoesterase |
| 0.514 | 3440 | Hemolysin-type calcium-binding region | 3438 | No predicted function |
| | 3441 | No predicted function | 3439 | Quinate/Shikimate 5-dehydrogenase |

Note:

Each row presents a consecutive pair of previously predicted operons. The score is generated by a Logistic Model Tree classifier. Each pair presented here has classifier score >0.5, and therefore should likely be merged into a single longer predicted operon. The numbers in the "Prior 1" and "Prior 2" columns are gene identifiers, truncated for formatting; prepend "CwatDRAFT_" to the numbers to generate the full identifier. In the "Gene Functions" columns, genes with unknown function are noted in bold; known function of other genes in a predicted operon can provide clues to the unknown function.

zero. A positive training set of operons for the classifier was collected by identifying all prior predicted operons in which at least one gene's expression exhibited diel variation. A negative training set for the classifier was generated by identifying consecutive genes where

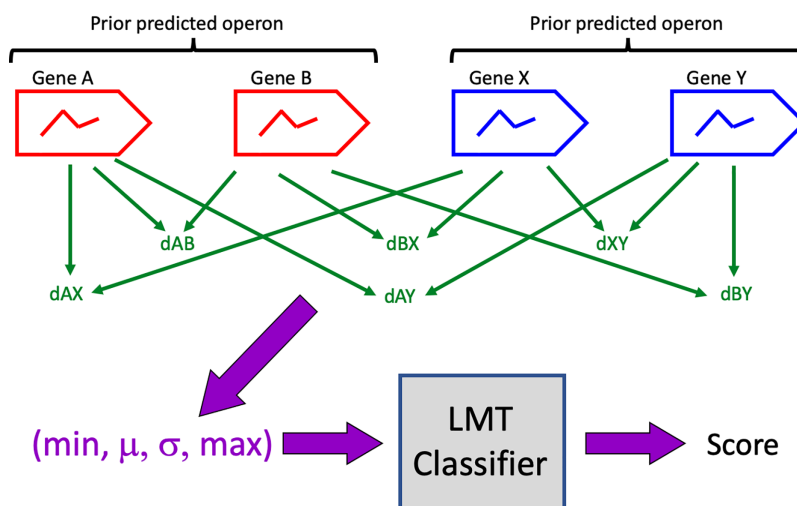


Figure 1 Classification algorithm. Prior predicted operons (red, blue) are candidates for merging if their genes are consecutive, if all genes are on the same DNA strand, and if at least one gene in each prior exhibits diel expression. [Full-size !\[\]\(b345a1c4255362eec3746050dd71ccac_img.jpg\) DOI: 10.7717/peerj.13259/fig-1](https://doi.org/10.7717/peerj.13259/fig-1)

at least one gene's expression exhibited diel variation, and where at least one gene is on each DNA strand. All genes of an operon must be on the same strand, to allow correct translation of the transcribed operon.

The classifier requires training and evaluation instances to be represented by vectors of numbers. For each prior in the training sets (and, later, for each merge candidate to be classified), the ABLIM distance between every pair of genes was computed; the instance was represented by a 4-vector consisting of the minimum, mean, standard deviation, and maximum of the ABLIM distances. Gene distances within priors were not considered, as they did not fit a Gaussian distributions and are therefore problematic. A total of 45 classifiers (Table S1) in the WEKA software suite (Hall et al., 2009; Frank, Hall & Witten, 2016) were evaluated on the positive and negative sets using five-fold cross-validation. The Logistic Model Tree (LMT) classifier gave the best accuracy on both the positive and negative data, and was therefore selected for the remainder of the study. The classifier was trained using all the positive and negative instances.

Pairs of prior predictions were identified as candidates for merging (Table 1) if there were no intervening genes, if all genes lay on the same DNA strand and in the same contig, and if each prior contained at least one gene whose expression exhibited diel variation. A 4-vector representation of each candidate was computed as described above, and the representations were evaluated on the trained LMT classifier to generate classification scores (Fig. 1). A candidate was accepted (*i.e.* all its genes are predicted to be in a single operon) if classifier score was >0.5 . Note that this score is not to be interpreted as a probability that the classification is correct.

Software was developed on Eclipse 2020-12 (4.18.0) in Java SE-15 (and is compatible with Java 1.8), using version 3.5 of the WEKA library and version 3.6.1 of the Apache Commons Mathematics Library.

RESULTS

The positive training set consists of the 1,195 operon predictions at <http://www.microbesonline.org/operons/>. The negative training set is listed in [Table S1](#). A total of 45 classifiers in the WEKA software were evaluated on the training data. The Logistical Model Tree (LMT) had the highest accuracy ([Table S1](#)).

A total of 63 pairs of prior operon predictions were identified as candidates for merging. Each prior consisted of two genes, at least one of which exhibited diel expression variation; all genes were on the same DNA strand and in the same contig, and there were no intervening genes between the priors. A total of 15 pairs of priors were classified as belonging to a common operon ([Table 1](#)).

DISCUSSION

Diel expression data was combined with prior operon predictions to compute 15 pairs of priors ([Table 1](#)) that appear to belong to common operons. It is recommended that each of these pairs be merged into a single prediction.

One reason for honing operon predictions is to gain insight into the function of unknown genes. When unknown genes share an operon with genes of known function, the known function can reasonably be hypothesized to relate to the unknown functions. In [Table 1](#), unknown genes are marked in bold. Six prior predictions include operons where no gene has known function; in all these cases, the present analysis predicts that the prior prediction should be merged with another prior containing at least one gene of known function. Predicted operon membership *per se* may not be strong enough evidence to infer gene function, but it can provide the basis for hypothesizing function, and the hypothesis can be strengthened by other evidence.

Each operon (training priors and merge candidates) was represented by a 4-vector consisting of the minimum, mean, standard deviation, and maximum of the ABLIM distances among all gene pairs in the operon. None of these statistics alone was sufficient for training an accurate classifier. The LMT classifier had the best accuracy among the 45 classifiers that were evaluated ([Table S1](#)). However, this does not imply that LMT should be used when analyzing other organisms. Future work on other organisms should repeat the classifier evaluation reported here, and should choose the best classifier for the organism at hand.

CONCLUSIONS

The work presented here demonstrates that machine learning analysis of diel expression studies can improve *in silico* predictions of operons. When a prior prediction is extended to include genes of unknown function, the function of the known genes in the prior might elucidate the function of the new unknown genes.

The approach presented here can be applied to other cyanobacteria for which diel studies and prior predicted operons are available. Since the method is based on similarity of diel signatures, best results should be expected from organisms whose genes exhibit strong and diverse diel variation. Organisms with weak diel variation can be expected to perform poorly, because the 4-vectors that describe operons to the classifier would all be

similar. Experiments with a diel study (Zinser *et al.*, 2009) of the minimal bacterium *Prochlorococcus marinus* produced poor results with the approach presented here, possibly because the circadian clock mechanism is simplified in *Prochlorococcus* (Holtzendorff *et al.*, 2008) and its diel genes fluctuate more weakly than those of *Crocospaera*.

ACKNOWLEDGEMENTS

The author is grateful to Jonathan Zehr, Josh Stuart, Irina Shilova, Rex Malmstrom, and Laurence Nedelec for valuable discussions.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by a Gordon and Betty Moore Marine Investigator grant and the Microbial Environmental Genomics Applications: Modeling, Experimentation, and Remote Sensing (MEGAMER) facility of the University of California, Santa Cruz. There was no additional external funding received for this study. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Gordon and Betty Moore Marine Investigator.
University of California, Santa Cruz.

Competing Interests

The author declares that he has no competing interests.

Author Contributions

- Philip Heller conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

Source code and instructions are available at GitHub: <https://github.com/PhilipHeller/Operons> (DOI 10.5281/zenodo.5759925).

The third-party data is available from MicrobesOnline Operon Predictions <http://www.microbesonline.org/operons/>. Contact Eric Alm (ejalm@mit.edu).

Supplemental Information

Supplemental information for this article can be found online at <http://dx.doi.org/10.7717/peerj.13259#supplemental-information>.

REFERENCES

- Bergersen FJ. 1962.** The effects of partial pressure of oxygen upon respiration and nitrogen fixation by soybean root nodules. *Journal of General Microbiology* **29**(1):113–125
DOI [10.1099/00221287-29-1-113](https://doi.org/10.1099/00221287-29-1-113).
- Chittapun S, Limbipichai S, Amnuaysin N, Boonkerd R, Charoensook M. 2018.** Effects of using cyanobacteria and fertilizer on growth and yield of rice, Pathum Thani I: a pot experiment. *Journal of Applied Phycology* **30**(1):79–85 DOI [10.1007/s10811-017-1138-y](https://doi.org/10.1007/s10811-017-1138-y).
- Dodd AN. 2005.** Plant circadian clocks increase photosynthesis, growth, survival, and competitive advantage. *Science* **309**(5734):630–633 DOI [10.1126/science.1115581](https://doi.org/10.1126/science.1115581).
- Farrokh P, Sheikhpour M, Kasaiean A, Asadi H, Bavandi R. 2019.** Cyanobacteria as an eco-friendly resource for biofuel production: a critical review. *Biotechnology Progress* **35**(5):21 DOI [10.1002/btpr.2835](https://doi.org/10.1002/btpr.2835).
- Fay P. 1992.** Oxygen relations of nitrogen fixation in cyanobacteria. *Microbiological Reviews* **56**(2):340–373 DOI [10.1128/mr.56.2.340-373.1992](https://doi.org/10.1128/mr.56.2.340-373.1992).
- Frank E, Hall M, Witten I. 2016.** The WEKA Workbench. In: *Data Mining: Practical Machine Learning Tools and Techniques*. Burlington: Morgan Kaufmann.
- Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. 2009.** The WEKA data mining software. *ACM SIGKDD Explorations Newsletter* **11**(1):10–18 DOI [10.1145/1656274.1656278](https://doi.org/10.1145/1656274.1656278).
- Heller P, Baiju B. 2018.** An improved distance metric for clustering gene expression time-series data. *American Journal of Advanced Research* **2**(1):14–17 DOI [10.5281/zenodo.1419822](https://doi.org/10.5281/zenodo.1419822).
- Holtzendorff J, Partensky F, Mella D, Lennon JF, Hess WR, Garczarek L. 2008.** Genome streamlining results in loss of robustness of the circadian clock in the marine cyanobacterium *Prochlorococcus marinus* PCC 9511. *Journal of Biological Rhythms* **23**:187–199 DOI [10.1177/0748730408316040](https://doi.org/10.1177/0748730408316040).
- Khan Z, Bhadouria P, Bisen P. 2005.** Nutritional and therapeutic potential of spirulina. *Current Pharmaceutical Biotechnology* **6**(5):373–379 DOI [10.2174/138920105774370607](https://doi.org/10.2174/138920105774370607).
- Landwehr N, Hall M, Frank E. 2005.** Logistic model trees. *Machine Learning* **59**(1–2):161–205 DOI [10.1007/s10994-005-0466-3](https://doi.org/10.1007/s10994-005-0466-3).
- Lercher MJ. 2003.** Coexpression of neighboring genes in *Caenorhabditis Elegans* is mostly due to operons and duplicate genes. *Genome Research* **13**(2):238–243 DOI [10.1101/gr.553803](https://doi.org/10.1101/gr.553803).
- Moreno-Hagelsieb G, Collado-Vides J. 2002.** A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics* **18**:S329–S336 DOI [10.1093/bioinformatics/18.suppl_1.s329](https://doi.org/10.1093/bioinformatics/18.suppl_1.s329).
- Morone J, Alfeus A, Vasconcelos V, Martins R. 2019.** Revealing the potential of cyanobacteria in cosmetics and cosmeceuticals—a new bioactive approach. *Algal Research* **41**(12):101541 DOI [10.1016/j.algal.2019.101541](https://doi.org/10.1016/j.algal.2019.101541).
- Price MN. 2005.** A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Research* **33**:880–892 DOI [10.1093/nar/gki232](https://doi.org/10.1093/nar/gki232).
- Renger G, Völker M, Eckert HJ, Fromme R, Hohm-Veit S, Gräber P. 1989.** On the mechanism of Photosystem II deterioration by UV-B irradiation. *Photochemistry and Photobiology* **49**:97–105 DOI [10.1111/j.1751-1097.1989.tb04083.x](https://doi.org/10.1111/j.1751-1097.1989.tb04083.x).
- Sabatti C. 2002.** Co-expression pattern from DNA microarray experiments as a tool for operon prediction. *Nucleic Acids Research* **30**(13):2886–2893 DOI [10.1093/nar/gkf388](https://doi.org/10.1093/nar/gkf388).
- Sakurai H, Masukawa H, Kitashima M, Inoue K. 2015.** How close we are to achieving commercially viable large-scale photobiological hydrogen production by cyanobacteria: a review of the biological aspects. *Life* **5**(1):997–1018 DOI [10.3390/life5010997](https://doi.org/10.3390/life5010997).
- Schopf J. 2002.** The fossil record: tracing the roots of the cyanobacterial lineage. In: Whitton B, Potts M, eds. *The Ecology of Cyanobacteria*. Alphen aan den Rijn: Kluwer, 13–35.

- Shi T, Ilikchyan I, Rabouille S, Zehr JP. 2010.** Genome-wide analysis of diel gene expression in the unicellular N₂-fixing cyanobacterium *Crocospaera watsonii* WH 8501. *The ISME Journal* 4(5):621–632 DOI 10.1038/ismej.2009.148.
- Singh H, Khattar JS, Ahluwalia AS. 2014.** Cyanobacteria and agricultural crops. *International Journal of Plant Research* 27:37 DOI 10.5958/j.2229-4473.27.1.008.
- Soni B, Trivedi U, Madamwar D. 2008.** A novel method of single step hydrophobic interaction chromatography for the purification of phycocyanin from *Phormidium fragile* and its characterization for antioxidant property. *Bioresource Technology* 99(1):188–194 DOI 10.1016/j.biortech.2006.11.010.
- Sumner M, Frank E, Hall M. 2005.** Speeding up logistic model tree induction. In: Jorge AM, Torgo L, Brazdil P, Camacho R, Gama J, eds. *Knowledge Discovery in Databases: PKDD 2005*. Vol. 3721. Berlin: Springer, 675–683.
- Tan LT. 2007.** Bioactive natural products from marine cyanobacteria for drug discovery. *Phytochemistry* 68(7):954–979 DOI 10.1016/j.phytochem.2007.01.012.
- Tang W, Li Z, Cassar N. 2019.** Machine learning estimates of global marine nitrogen fixation. *Journal of Geophysical Research: Biogeosciences* 124(3):717–730 DOI 10.1029/2018JG004828.
- Tomitani A, Knoll AH, Cavanaugh CM, Ohno T. 2006.** The evolutionary diversification of cyanobacteria: molecular-phylogenetic and paleontological perspectives. *Proceedings of the National Academy of Sciences of the United States of America* 103(14):5442–5447 DOI 10.1073/pnas.0600999103.
- Tréguer P, Bowler C, Moriceau B, Dutkiewicz S, Gehlen M, Aumont O, Bittner L, Dugdale R, Finkel Z, Iudicone D, Jahn O, Guidi L, Lasbleiz M, Leblanc K, Levy M, Pondaven P. 2018.** Influence of diatom diversity on the ocean biological carbon pump. *Nature Geoscience* 11(1):27–37 DOI 10.1038/s41561-017-0028-x.
- Tuit C, Waterbury J, Ravizza G. 2004.** Diel variation of molybdenum and iron in marine diazotrophic cyanobacteria. *Limnology and Oceanography* 49(4):978–990 DOI 10.4319/lo.2004.49.4.0978.
- Wang W-L, Moore JK, Martiny AC, Primeau FW. 2019.** Convergent estimates of marine nitrogen fixation. *Nature* 566(7743):205–211 DOI 10.1038/s41586-019-0911-2.
- Westover BP, Buhler JD, Sonnenburg JL, Gordon JI. 2005.** Operon prediction without a training set. *Bioinformatics* 21:880–888 DOI 10.1093/bioinformatics/bti123.
- Yao DCI, Brune DC, Vavilin D, Vermaas WFJ. 2012.** Photosystem II component lifetimes in the cyanobacterium *Synechocystis* sp. Strain PCC 6803: small cab-like proteins stabilize biosynthesis intermediates and affect early steps in chlorophyll synthesis. *Journal of Biological Chemistry* 287:682–692 DOI 10.1074/jbc.M111.320994.
- Zanchett G, Oliveira-Filho E. 2013.** Cyanobacteria and cyanotoxins: from impacts on aquatic ecosystems and human health to anticarcinogenic effects. *Toxins* 5(10):1896–1917 DOI 10.3390/toxins5101896.
- Zehr JP. 2011.** Nitrogen fixation by marine cyanobacteria. *Trends in Microbiology* 19(4):162–173 DOI 10.1016/j.tim.2010.12.004.
- Zheng Y, Szustakowski JD, Fortnow L, Roberts RJ, Kasif S. 2002.** Computational identification of operons in microbial genomes. *Genome Research* 12(8):1221–1230 DOI 10.1101/gr.200602.
- Zinser ER, Lindell D, Johnson ZI, Futschik ME, Steglich C, Coleman ML, Wright MA, Rector T, Steen R, McNulty N, Thompson LR, Chisholm SW. 2009.** Choreography of the transcriptome, photophysiology, and cell cycle of a minimal photoautotroph, prochlorococcus. *PLOS ONE* 4(1):e5135 DOI 10.1371/journal.pone.0005135.