Diel gene expression improves software prediction of cyanobacterial operons

Philip Heller Corresp. 1

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

Corresponding Author: Philip Heller
Email address: philip.heller@sjsu.edu

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 *Crocosphaera watsonii*. 22 operon enhancements are proposed.

Diel gene expression improves software prediction of cyanobacterial operons

Philip Heller[1]

[1] Department of Computer Science, San Jose State University, San Jose, CA, USA

Corresponding Author:
Philip Heller[1] 1 Washington Sq., San Jose, CA 95192, USA
Email address: philip.heller@sjsu.edu

Abstract

Add your abstract here. 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 *Crocosphaera watsonii*. 22 operon enhancements are proposed.

Introduction

Photosynthesizing bacteria (Phylum *Cyanobacteria*) are significant participants in global biogeochemical cycles. They arose on Earth 3.5 billion years ago[1], and had oxygenated the atmosphere by 2.5 billion years ago[2]. Cyanobacteria participate in the ocean biological carbon pump[3], 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[4] [5]. Cyanobacteria are used to produce medicines[6] [7] [8], biofuels[9] [10], fertilizers[11] [12], cosmetics[13], and food[14].

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[15]. In particular, identification of operons (consecutive genes 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[16 17], supports interpretation of transcriptome experiments[18], and can guide annotation of hypothetical genes. The expense of wetlab operon discovery has prompted the development of algorithms for predicting operons from assembled genomes[16 18 19]; predictions from one of these algorithms[19] for 1336 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-hour 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 hours[20 21], approximately coincides with available light[22]. Since oxygen disables nitrogenase (the enzyme responsible for nitrogen fixation), diazotrophic cyanobacteria segregate nitrogenase from the oxygen evolved by photosynthesis[23 24]; segregation is sometimes temporal, with nitrogenase component proteins produced hours out of phase from photosystem II proteins[25]. Diel cycling, defined as a transcript abundance change of at least 2x over 24 hours, has been observed in 79% of genes of the diazotrophic cyanobacterium *Crocosphaera watsonii*[26]. Since genes in an operon are expected to have similar expression signatures[27 28], 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[29 30] (LMT) - to determine when predicted operons in *Crocosphaera* should be merged. A common metric for quantifying expression similarity is Pearson's Correlation Coefficient (PCC); however, our earlier work[31] 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[31], 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 *Crocosphaera watsonii* genome. 48 kinds of classifier (Supplemental Table 1) 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. 22 pairs of candidate operon predictions are recommended for merging (Table 1).

Materials & Methods

Computed operon predictions (hereafter the "prior predictions") for strain *Crocosphaera watsonii* were downloaded from http://www.microbesonline.org/operons/. Log-expression measurements for 4,407 *Crocosphaera* genes with 8 timepoints were retrieved from a study by Shi et al.[26] For each gene, log-expressions were normalized to a mean of zero. A positive training set of operons for the classifier was collected by identifying all prior predicted operons in which at least 1 gene's expression exhibited diel variation. A negative training set for the classifier was generated by identifying consecutive genes where at least 1 gene's expression exhibited diel variation, and where each DNA strand is represented. (Since operons are transcribed as a single unit, and transcription is restricted to a single strand, these sets of genes cannot be operons.)

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. 48 classifiers (Supplemental Table 1) in the WEKA software suite[32] [33] were evaluated on the positive and negative sets using 5-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 (Supplemental Table 2) 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 1 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 (Figure 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.

To estimate the accuracy of the classifier's predictions, each negative training example in turn was censored from the training set; the model was then re-trained on the remaining data, and the censored example's classification score was computed. A Gaussian distribution was computed for the classification scores thus generated. Given a candidate with score s, the cumulative probability of scores >= s is an estimate of the probability of erroneously accepting the candidate. Table 1 lists the accepted predictions, with their classifier scores and estimated error probabilities.

Results

The positive training set consists of the 1195 operon predictions at http://www.microbesonline.org/operons/. The negative training set is listed in Supplemental Table 1. 48 classifiers in the WEKA software were evaluated on the training data. The Logistical Model Tree (LMT) had the highest accuracy (Supplemental Table 1).

79 pairs of prior operon predictions were identified as candidates for merging. Each prior consisted of 2 genes, at least 1 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 2 priors. 22 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 22 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 underlined bold. 8 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 1 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 44 classifiers that were evaluated (Supplemental Table 1). 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.

The false-positive probability (column “P(false +)” in Supplemental Table 2) is a rough estimate. It has much in common with a p-value: the null hypothesis is that the prior operons should not be merged; the alternative hypothesis is that they should be merged; the statistic is the cumulative probability of the null hypothesis when a score is at least as strong as the score at hand. However, the cumulative probability is based on a negative training set of non-operons which is specific but not sensitive. No members of the negative set can possibly be operons, because both DNA strands are present. However no same-strand non-operons are present in the negative training set, because these are difficult to ascertain. Thus there is a bias in the negative set, and the resulting P(false +) values should not be viewed as rigorous.

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[34] of the minimal bacterium *Prochlorococcus marinus* produced poor results with the approach presented here, possibly because the circadian clock mechanism is simplified in *Prochlorococcus*[35] and its diel genes fluctuate more weakly than those of *Crocosphaera.*

Acknowledgements

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

References

1. Schopf, J. In: Whitton B, Potts M (eds) The fossil record: tracing the roots of the cyanobacterial lineage. in *The Ecology of Cyanobacteria* 13–35 (Kluwer, 2002).

2. Tomitani, A., Knoll, A. H., Cavanaugh, C. M. & Ohno, T. The evolutionary diversification of cyanobacteria: Molecular-phylogenetic and paleontological perspectives. *Proceedings of the National Academy of Sciences* **103**, 5442–5447 (2006).

3. Tréguer, P. *et al.* Influence of diatom diversity on the ocean biological carbon pump. *Nature Geosci* **11**, 27–37 (2018).

4. Wang, W.-L., Moore, J. K., Martiny, A. C. & Primeau, F. W. Convergent estimates of marine nitrogen fixation. *Nature* **566**, 205–211 (2019).

5. Tang, W., Li, Z. & Cassar, N. Machine Learning Estimates of Global Marine Nitrogen Fixation. *Journal of Geophysical Research: Biogeosciences* **124**, 717–730 (2019).

6. Tan, L. T. Bioactive natural products from marine cyanobacteria for drug discovery. *Phytochemistry* **68**, 954–979 (2007).

7. Soni, B., Trivedi, U. & Madamwar, D. 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**, 188–194 (2008).

8. Zanchett, G. & Oliveira-Filho, E. Cyanobacteria and Cyanotoxins: From Impacts on Aquatic Ecosystems and Human Health to Anticarcinogenic Effects. *Toxins* **5**, 1896–1917 (2013).

9. Sakurai, H., Masukawa, H., Kitashima, M. & Inoue, K. How Close We Are to Achieving Commercially Viable Large-Scale Photobiological Hydrogen Production by Cyanobacteria: A Review of the Biological Aspects. *Life* **5**, 997–1018 (2015).

10. Farrokh, P., Sheikhpour, M., Kasaeian, A., Asadi, H. & Bavandi, R. Cyanobacteria as an eco-friendly resource for biofuel production: A critical review. *Biotechnol. Prog.* **35**, (2019).

11. Singh, H., Khattar, J. S. & Ahluwalia, A. S. Cyanobacteria and agricultural crops. *Vegetos- An Inter. Jour. of Plnt. Rese.* **27**, 37 (2014).

12. Chittapun, S., Limbipichai, S., Amnuaysin, N., Boonkerd, R. & Charoensook, M. Effects of using cyanobacteria and fertilizer on growth and yield of rice, Pathum Thani I: a pot experiment. *J Appl Phycol* **30**, 79–85 (2018).

13. Morone, J., Alfeus, A., Vasconcelos, V. & Martins, R. Revealing the potential of cyanobacteria in cosmetics and cosmeceuticals — A new bioactive approach. *Algal Research* **41**, 101541 (2019).

14. Khan, Z., Bhadouria, P. & Bisen, P. Nutritional and Therapeutic Potential of Spirulina. *CPB* **6**, 373–379 (2005).

15. Zehr, J. P. Nitrogen fixation by marine cyanobacteria. *Trends in Microbiology* **19**, 162–173 (2011).

16. Zheng, Y., Szustakowski, J. D., Fortnow, L., Roberts, R. J. & Kasif, S. Computational Identification of Operons in Microbial Genomes. *Genome Research* **12**, 1221–1230 (2002).

17. Westover, B. P., Buhler, J. D., Sonnenburg, J. L. & Gordon, J. I. Operon prediction without a training set. *Bioinformatics* **21**, 880–888 (2005).

18. Moreno-Hagelsieb, G. & Collado-Vides, J. A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics* **18**, S329–S336 (2002).

19. Price, M. N. A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Research* **33**, 880–892 (2005).

20. Yao, D. C. I., Brune, D. C., Vavilin, D. & Vermaas, W. F. J. 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 (2012).

21. Renger, G. *et al.* ON THE MECHANISM OF PHOTOSYSTEM II DETERIORATION BY UV-B IRRADIATION. *Photochemistry and Photobiology* **49**, 97–105 (1989).

22. Dodd, A. N. Plant Circadian Clocks Increase Photosynthesis, Growth, Survival, and Competitive Advantage. *Science* **309**, 630–633 (2005).

23. Bergersen, F. J. The Effects of Partial Pressure of Oxygen upon Respiration and Nitrogen Fixation by Soybean Root Nodules. *Journal of General Microbiology* **29**, 113–125 (1962).

24. Fay, P. Oxygen Relations of Nitrogen Fixation in Cyanobacteria. *MICROBIOL. REV.* **56**, 34 (1992).

25. Tuit, C., Waterbury, J. & Ravizza, G. Diel variation of molybdenum and iron in marine diazotrophic cyanobacteria. *Limnology and Oceanography* **49**, 978–990 (2004).

26. Shi, T., Ilikchyan, I., Rabouille, S. & Zehr, J. P. Genome-wide analysis of diel gene expression in the unicellular N2-fixing cyanobacterium Crocosphaera watsonii WH 8501. *The ISME Journal* **4**, 621–632 (2010).

27. Sabatti, C. Co-expression pattern from DNA microarray experiments as a tool for operon prediction. *Nucleic Acids Research* **30**, 2886–2893 (2002).

28. Lercher, M. J. Coexpression of Neighboring Genes in Caenorhabditis Elegans Is Mostly Due to Operons and Duplicate Genes. *Genome Research* **13**, 238–243 (2003).

29. Landwehr, N., Hall, M. & Frank, E. Logistic Model Trees. *Mach Learn* **59**, 161–205 (2005).

30. Sumner, M., Frank, E. & Hall, M. Speeding Up Logistic Model Tree Induction. in *Knowledge Discovery in Databases: PKDD 2005* (eds. Jorge, A. M., Torgo, L., Brazdil, P., Camacho, R. & Gama, J.) vol. 3721 675–683 (Springer Berlin Heidelberg, 2005).

31. Heller, P. & Baiju, Bharath. An Improved Distance Metric For Clustering Gene Expression Time-Series Data. *Zenodo* (2018) doi:10.5281/zenodo.1419822.

32. Hall, M. *et al.* The WEKA Data Mining Software: An Update. **11**, 9 (2009).

33. Frank, E., Hall, M. & Witten, I. The WEKA Workbench. in *Data Mining: Practical Machine Learning Tools and Techniques* (Morgan Kaufmann, 2016).

34. Zinser, E. R. *et al.* Choreography of the Transcriptome, Photophysiology, and Cell Cycle of a Minimal Photoautotroph, Prochlorococcus. *PLoS ONE* **4**, e5135 (2009).

35. Holtzendorff, J. *et al.* Genome Streamlining Results in Loss of Robustness of the Circadian Clock in the Marine Cyanobacterium *Prochlorococcus marinus* PCC 9511. *J Biol Rhythms* **23**, 187–199 (2008).

Table 1(on next page)

The 48 classifiers evaluated in this study.

48 classifiers, evaluated by 5-fold cross validation on the *Crocosphaera* time-series data, in descending order of accuracy. Best accuracy was achieved by the Logistic Model Tree (LMT) classifier, which was selected for this study.

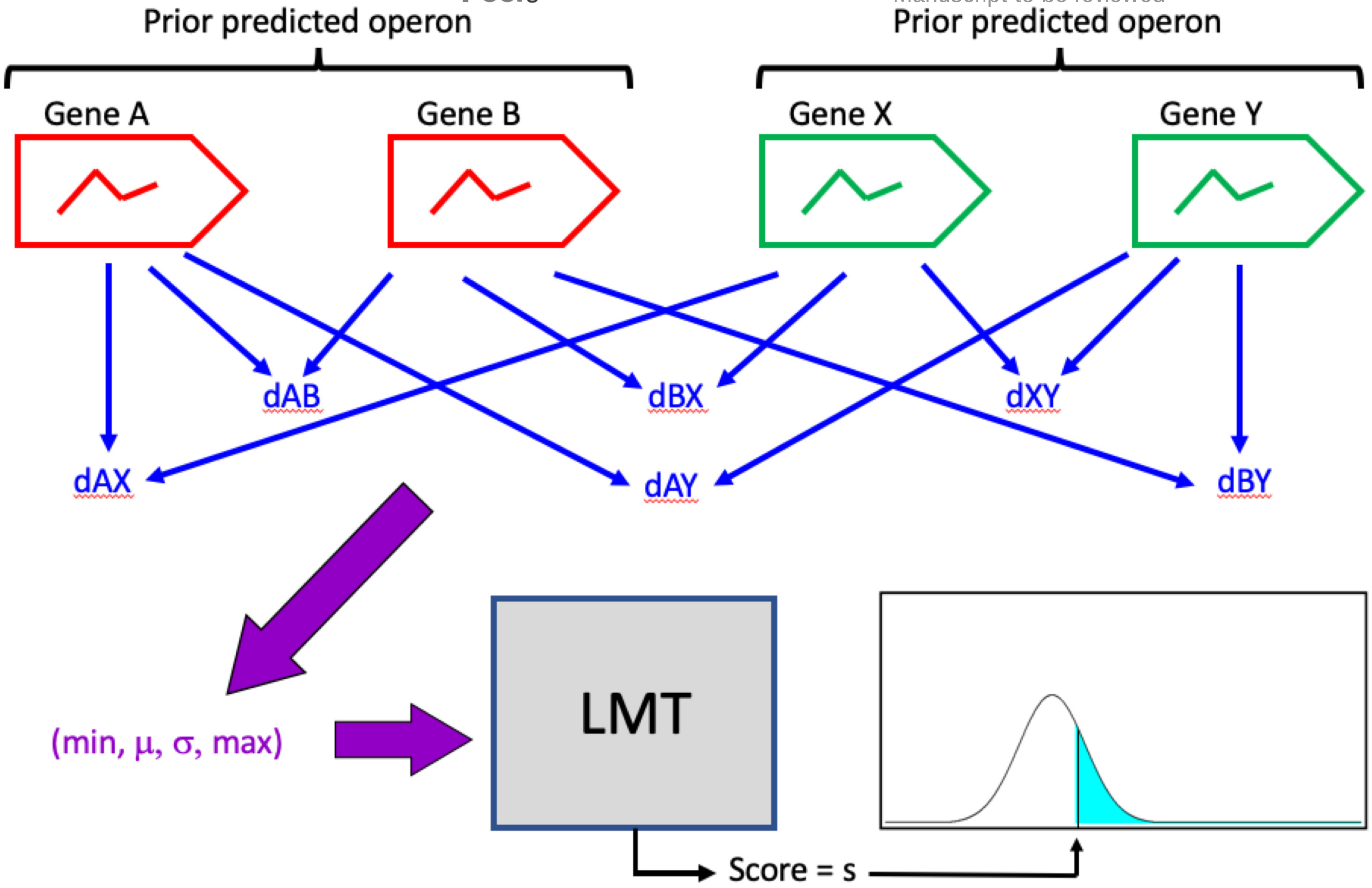
Classifier	Accuracy
trees.LMT	95.28
functions.MultilayerPerceptron	95.03
functions.Logistic	92.8
meta.MultiClassClassifier	92.8
meta.RandomizableFilteredClassifier	91.89
functions.SimpleLogistic	91.56
lazy.IBk	91.31
trees.RandomForest	91.14
lazy.KStar	90.89
meta.Bagging	89.82
meta.RandomCommittee	89.32
trees.RandomTree	88.91
functions.SGD	88.49
meta.MultiClassClassifierUpdateable	88.49
trees.REPTree	88.49
trees.J48	88.25
rules.JRip	87.83
meta.RandomSubSpace	87.67
functions.VotedPerceptron	87.33
rules.DecisionTable	87.17
meta.LogitBoost	86.51
meta.FilteredClassifier	86.26
meta.IterativeClassifierOptimizer	86.09
rules.PART	85.6
meta.AdaBoostM1	85.51
functions.SMO	85.1
bayes.BayesNet	84.44
trees.HoeffdingTree	83.53
bayes.NaiveBayesMultinomial	83.44
bayes.NaiveBayesMultinomialUpdateable	83.44
bayes.NaiveBayes	82.7
bayes.NaiveBayesUpdateable	82.7
meta.AttributeSelectedClassifier	82.12
lazy.LWL	81.71
trees.DecisionStump	81.71
rules.OneR	81.29
bayes.NaiveBayesMultinomialText	66.14
functions.SGDText	66.14
meta.CVParameterSelection	66.14
meta.MultiScheme	66.14
meta.Stacking	66.14
meta.Vote	66.14
meta.WeightedInstancesHandlerWrapper	66.14
misc.InputMappedClassifier	66.14
rules.ZeroR	66.14

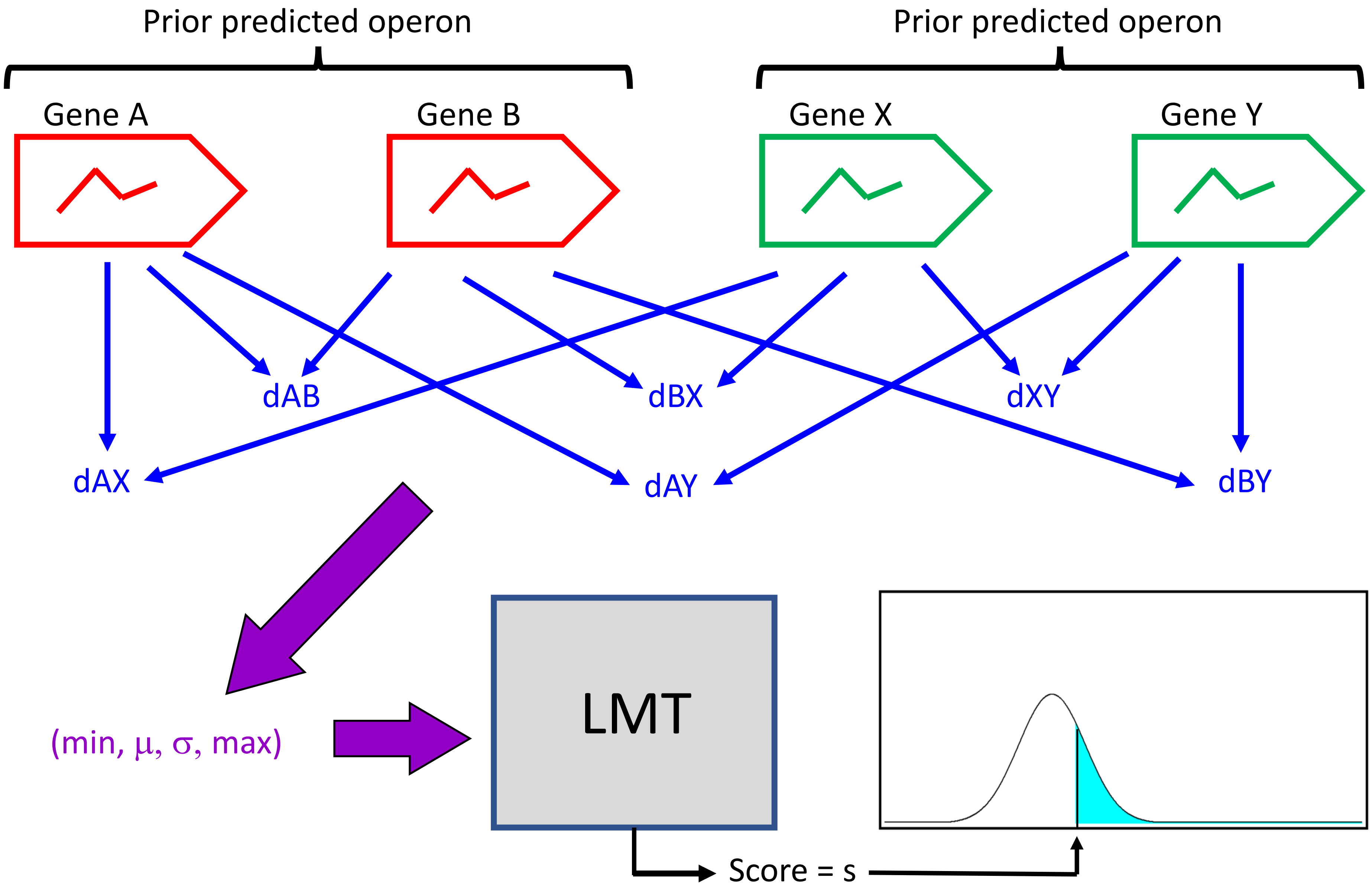
Supplemental Table 1 – 48 Classifiers evaluated by 5-fold cross validation, in descending order of accuracy. The LMT (Logistic Model Tree) classifier was chosen for this study.

Figure 1(on next page)

Method for accepting/rejecting proposed merger of 2 prior predicted operons (red and green).

Both priors must lie on the same DNA strand, and each must contain at least 1 gene with diel variation. All pairwise ABLIM distances among the 4 genes are computed (blue). A 4-vector comprising the minimum, mean, standard deviation, and maximum of the ABLIM distances is computed (purple) and submitted to the LMT classifier. The classifier produces a score s. Gaussian distributions over scores of positive (upper curve) and negative (lower curve) are used to compute, respectively, the false negative and false positive probabilities.


Prior predicted operon
Prior predicted operon
Gene A
Gene B
Gene X
Gene Y
dAB
dBX
dXY
dAX
dAY
dBY
(min, μ, σ, max)
LMT
Score = s


Prior predicted operon
Prior predicted operon
Gene A
Gene B
Gene X
Gene Y
dAB
dBX
dXY
dAX
dAY
dBY
(min, μ, σ, max)
LMT
Score = s