Understanding *Cyanothece* 51142’s clock module using Mutual Information

The cyanobacterial circadian clock mechanism has been experimentally validated in *Synechococcus Elongatus* sp. 7942 (1). The core clock comprises of 3 proteins, KaiA, KaiB and KaiC. The clock proteins receive signals from the external environment through elements of the input pathway such as CikA which is a histidine kinase. The output pathway is mediated through the sensory histidine kinase sasA and a transcriptional regulator RpaA. In *Cyanothece* sp. 51142 however, there are multiple copies of the Kai genes, the two-component kinases that interact with the clock genes and the two component sensors and regulators that are supposed to be a part of the output pathway (2). The circadian clock signaling network in *Cyanothece* is thus expected to be more complicated than *Synechococcus* clock network described above. We intend to enhance our knowledge of *Cyanothece*’s signaling network by selecting the two component sensors and regulators and studying the correlation of their expression profile with the clock genes. This will also help guide experimental efforts in the Pakrasi lab by identifying potential clock genes to knockout.

We used Mutual Information as a correlation metric because of its ability to identify non-linear interactions as opposed to Pearson correlation which can only capture linear relationships. In addition to finding sensors and regulators that are highly correlated with the clock genes, we also infer whether a sensor-regulator pair belongs to the same two-component system. Since they are a part of the same system, it follows that their expression profiles should also be very highly correlated. Thus, we can gain more insights into the *Cyanothece* clock module by using Mutual Information.

Table of Contents

[Mutual Information 2](#_Toc37866512)

[MicroArray Dataset 2](#_Toc37866513)

[Analysis 3](#_Toc37866514)

[Stockel Dataset 3](#_Toc37866515)

[Toepel Dataset 5](#_Toc37866516)

[Comparing the two datasets 6](#_Toc37866517)

[Finding the most active set of KaiABC combination 8](#_Toc37866518)

[Stockel Dataset 9](#_Toc37866519)

[Toepel Dataset 9](#_Toc37866520)

[Case Study of some genes of interest 9](#_Toc37866521)

[cce\_1983/aphA 9](#_Toc37866522)

[cce\_0678 9](#_Toc37866523)

[cce\_0298/rpaA 10](#_Toc37866524)

[cce\_1751/sasA 10](#_Toc37866525)

[cce\_4002/rpaB 10](#_Toc37866526)

[cce\_4751/cikA 10](#_Toc37866527)

[cce\_0888/nblS 10](#_Toc37866528)

[Conclusion 10](#_Toc37866529)

[Appendix A 12](#_Toc37866530)

[Appendix B 13](#_Toc37866531)

[Stockel Dataset 13](#_Toc37866532)

[cce\_0678 13](#_Toc37866533)

[rpaA 14](#_Toc37866534)

[sasA 15](#_Toc37866535)

[rpaB 17](#_Toc37866536)

[Toepel Dataset 19](#_Toc37866537)

[cce\_0678 19](#_Toc37866538)

[rpaA 20](#_Toc37866539)

[sasA 21](#_Toc37866540)

[rpaB 22](#_Toc37866541)

[Appendix C 25](#_Toc37866542)

[Stockel Dataset 25](#_Toc37866543)

[Toepel Dataset 26](#_Toc37866544)

[References 27](#_Toc37866545)

# Mutual Information

If we consider two random variables X and Y, the Mutual Information (I) between them is defined as

where H(.) is the Shannon Entropy. For continuous variables, the idea is to estimate H(.) from the average distance to the k nearest neighbors (3). MI is thus estimated as

,

where N is the size of the dataset, k is the number of nearest neighbors and psi(x) is the digamma function, and <...> denotes averages of and and over all realizations of the random samples. are the number of points in the region and , are the edge lengths of the smallest rectangle around point containing nearest neighbors.

# MicroArray Dataset

The microarray expression datasets were obtained from the ArrayExpress database submitted by Stockel et. al. (4) and Toepel et. al. (5). The analysis was initially carried on in the Stockel et. al. dataset and the Toepel et. al. dataset was subsequently used to verify the results.

# Analysis

## Stockel Dataset

The expression profiles of the genes were first filtered to include only those genes that are annotated as circadian clock protein or two-component sensors and regulators. Gene annotations were obtained from the genomic database submitted by Welsh et. al. (6). Using the expression profiles of the selected genes, the mutual information between the individual clock genes and the (putative) sensors and regulators were obtained. To visualize the interactions, interaction matrices were created as shown below. The interaction of the clock genes with a two-component sensor, regulator, and a hybrid two-component sensor & regulator is given in Figure 1, Figure 2 and Figure 3 respectively (split into three for enhancing readability).

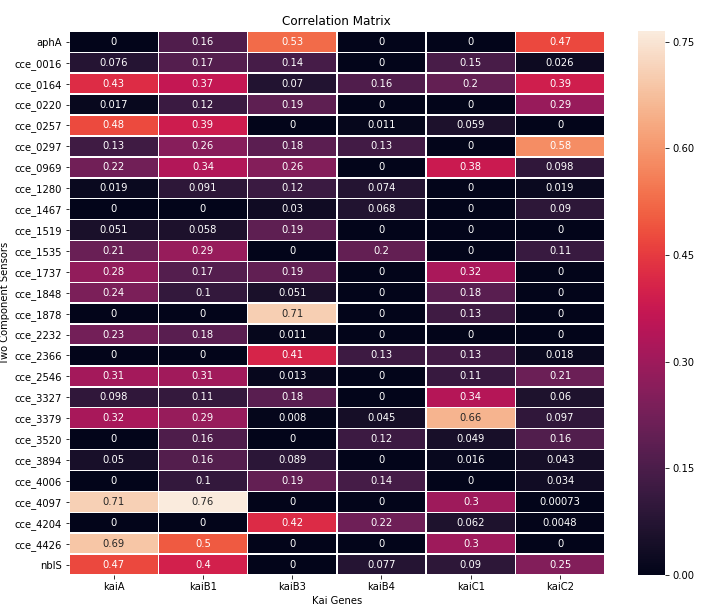


Figure 1: Interaction matrix of clock genes with all two-component systems present in *Cyanothece*. The x-axis contains the Kai clock genes and the y-axis contains the functionally annotated two-component sensors in *Cyanothece*. Each cell is colored according to the mutual information score between the corresponding Kai gene and a two-component sensor gene.

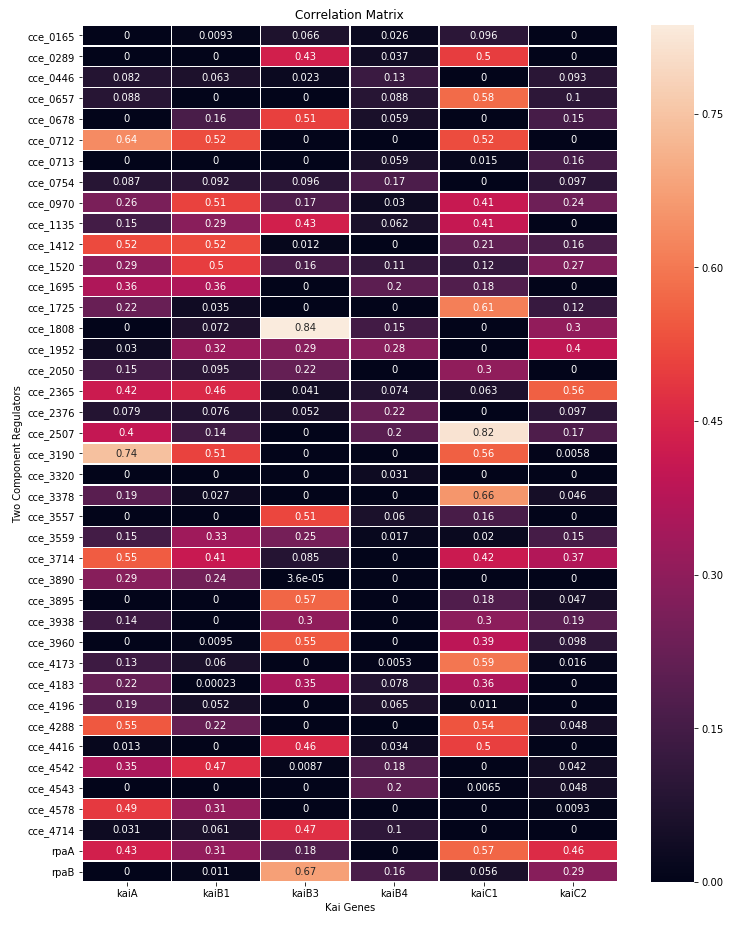


Figure 2: Interaction matrix of clock genes with all two-component systems present in *Cyanothece*. The x-axis contains the Kai clock genes and the y-axis contains the functionally annotated two-component sensors in *Cyanothece*. Each cell is colored according to the mutual information score between the corresponding Kai gene and a two-component sensor gene.

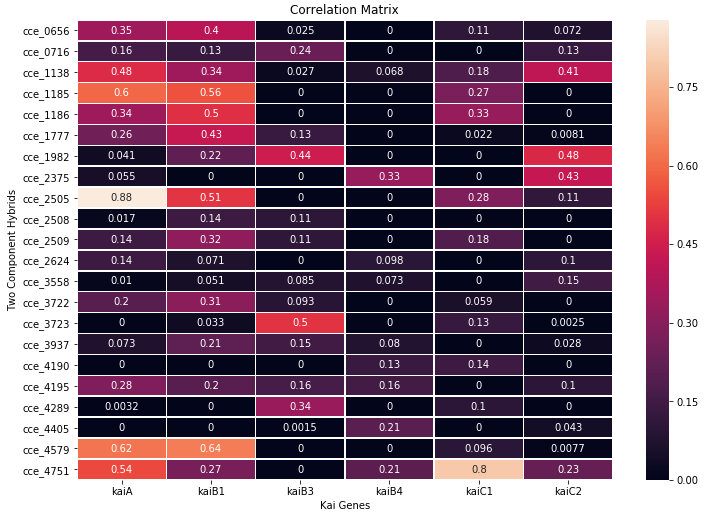


Figure 3: Interaction matrix of clock genes with all two-component systems present in *Cyanothece*. The x-axis contains the Kai clock genes and the y-axis contains the functionally annotated two-component sensors in *Cyanothece*. Each cell is colored according to the mutual information score between the corresponding Kai gene and a two-component sensor gene.

The interaction matrix between the clock genes inferred using mutual information is also presented below.

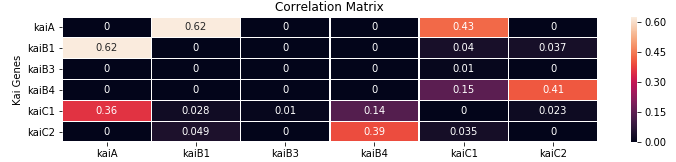


Figure 4: Interaction matrix between all clock genes present in *Cyanothece*. Each cell is colored according to the mutual information score between the corresponding Kai genes.

## Toepel Dataset

The expression profile of the genes in the Toepel Dataset were used to verify the results obtained from the Stockel dataset. The interaction of the clock proteins with the two-component sensors, regulators and the two-component hybrid sensor & regulator is given in Figure 5, Figure 6 and Figure 7 respectively.

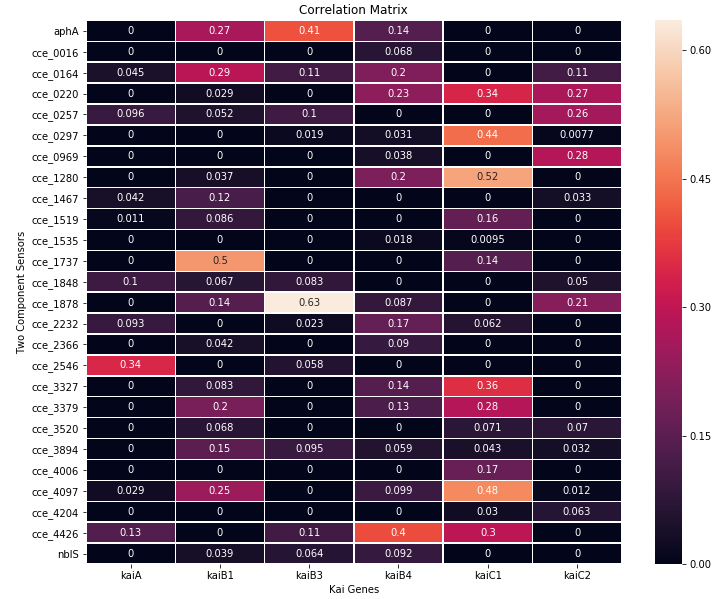


Figure 5: Interaction matrix of clock genes with all two-component systems present in *Cyanothece*. The x-axis contains the Kai clock genes and the y-axis contains the functionally annotated two-component sensors in *Cyanothece*. Each cell is colored according to the mutual information score between the corresponding Kai gene and a two-component sensor gene.

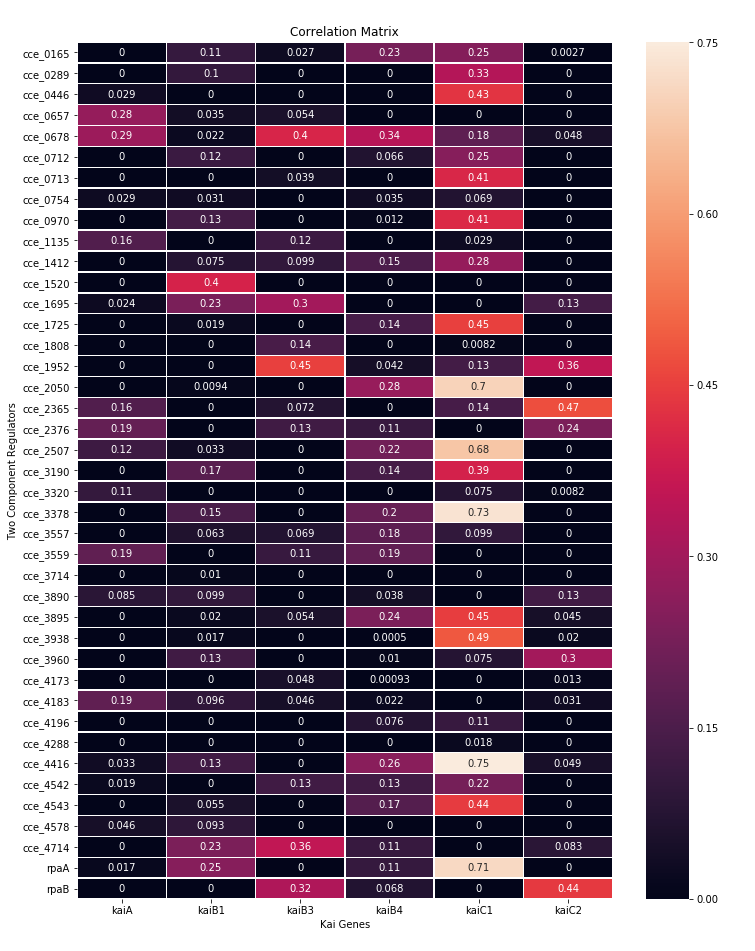


Figure 6: Interaction matrix of clock genes with all two-component systems present in *Cyanothece*. The x-axis contains the Kai clock genes and the y-axis contains the functionally annotated two-component sensors in *Cyanothece*. Each cell is colored according to the mutual information score between the corresponding Kai gene and a two-component sensor gene.

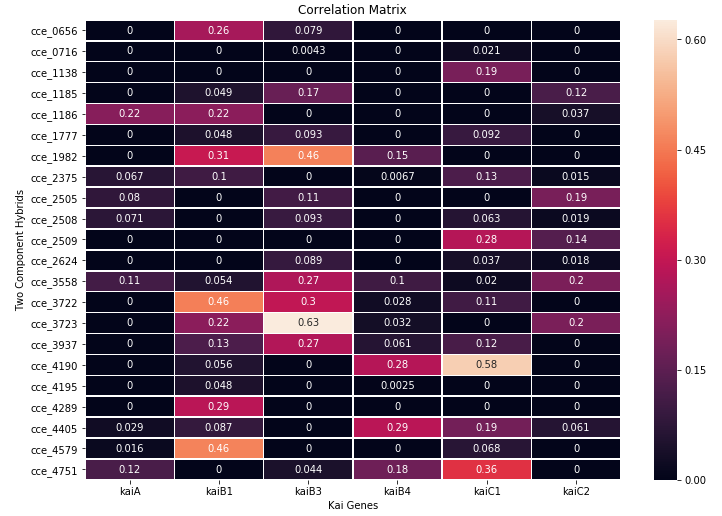


Figure 7: Interaction matrix of clock genes with all two-component systems present in *Cyanothece*. The x-axis contains the Kai clock genes and the y-axis contains the functionally annotated two-component sensors in *Cyanothece*. Each cell is colored according to the mutual information score between the corresponding Kai gene and a two-component sensor gene.

The correlation matrix between the clock genes is also presented below.

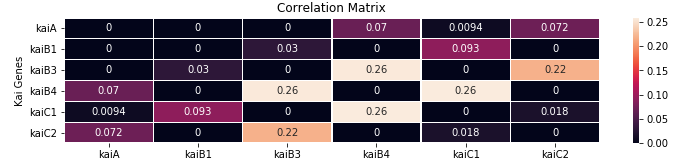


Figure 8: Interaction matrix between all clock genes present in *Cyanothece*. Each cell is colored according to the mutual information score between the corresponding Kai genes.

## Comparing the two datasets

* Among the sensors, genes aphA, cce\_1878, cce\_2546, cce\_3327, cce\_3379 and cce\_4426 show consistent mutual information scores with some of the Kai genes.
* Among the regulators, genes cce\_0678, cce\_0970, cce\_1725, cce\_1808, cce\_1952, cce\_2050, cce\_2365, cce\_2507, cce\_3190, cce\_3378, cce\_3938, cce\_4416, cce\_4714, rpaA and rpaB show consistent mutual information scores with some of the Kai genes.
* Among all hybrid sensors and regulators, genes cce\_0656, cce\_1138, cce\_1186, cce\_1983, cce\_2375, cce\_3558, cce\_3722, cce\_3723, cce\_3937, cce\_4190, cce\_4579, cce\_4751 show consistent mutual information scores with a subset of the Kai genes.
* The interaction among the Kai genes themselves lacks consistency, which may be due to fact that the two microarray datasets collected samples at different time points during a light/dark cycle.

# Finding the most active set of KaiABC combination

Since clock genes have multiple copies, the objective of this analysis was to find the KaiA, KaiB and KaiC combinations that are correlated to the maximum number of common sensors and regulators.

## Stockel Dataset

The table below (Table 1) presents the KaiABC combinations and the number of common sensors, regulators and hybrid sensors & regulators that they are interact with. It is arranged in descending order of the total number of common sensors, regulators and hybrids that the given KaiABC combination interacts with. Individual copies of the three Kai genes were grouped and ones with the maximum number of correlations with a common sensor and/or regulator were determined.

Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Kai Combination** | **No. of sensors** | **No. of regulators** | **No. of hybrids** | **Total** |
| kaiA-B1-C1 | 13 | 19 | 19 | 51 |
| kaiA-B1-C2 | 13 | 19 | 19 | 51 |
| kaiA-B3-C2 | 10 | 13 | 13 | 36 |
| kaiA-B3-C1 | 9 | 12 | 12 | 33 |
| kaiA-B4-C2 | 6 | 12 | 12 | 30 |
| kaiA-B4-C1 | 4 | 12 | 12 | 28 |

## Toepel Dataset

A similar table as above (Table 2) is presented below using the Toepel Dataset.

Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Kai Combination** | **No. of sensors** | **No. of regulators** | **No. of hybrids** | **Total** |
| kaiA-B1-C2 | 5 | 5 | 5 | 15 |
| kaiA-B4-C1 | 3 | 6 | 6 | 15 |
| kaiA-B3-C2 | 3 | 5 | 5 | 13 |
| kaiA-B1-C1 | 2 | 5 | 5 | 12 |
| kaiA-B4-C2 | 2 | 5 | 5 | 12 |
| kaiA-B3-C1 | 2 | 4 | 4 | 10 |

# Case Study of selected genes of interest

Here we present case studies of thoses genes that were either referenced in our previous report, proposed to be a part of the signaling network in other cyanobacteria, or were identified from this study. The complete list of interactions of cce\_0678, rpaA, rpaB and sasA is given in Appendix B. For a gene classified as a sensor, interactions with the regulators are given in the list and vice versa.

## cce\_1983/aphA

Cce\_1983 has been identified to have a very high mutual information with multiple clock genes in the current study using both Stockel and Toepel dataset. This histidine kinase has been functionally annotated as a probable phytochrome A in the *Cyanothece* Genomic Database developed by Welsh et. al. (6). Moreover, it also shares high mutual information score with the other genes of interest presented in this case study.

## cce\_0678

Cce\_0678 was identified in our previous analysis to be highly correlated with the RubisCo genes. In this study, it was found to have a high mutual information score with a number of sensors genes such as cce\_1983, cce\_0220, cce\_0164, cce\_2232. While cce\_1983 has already been discussed above, cce\_0220, cce\_0164 and cce\_2232 have a very high sequence similarity with the circadian input kinase cikA of *Synechococcus* 7942. Corresponding percent identity and E-values found using a protein-protein BLAST search for all these genes are given in Appendix A.

## cce\_0298/rpaA

Cce\_0298/rpaA has been shown to be a master regulator in *Synechococcus* (as stated in our previous report on Gene Regulatory Network of *Cyanothece*). This was corroborated in the present study as well, wherein it was found to interact with multiple clock genes, sensors and regulators. However, while in *Synechococcus*, sasA and rpaA belong to the two-component system, in *Cyanothece*, cce\_1751 (which has the highest sequence similarity to sasA from *Synechococcus*) was not found to interact with cce\_0298. However, cce\_1751 has a very high mutual information score with rpaB (cce\_4002). Therefore, our analysis predicts that in Cyanothece sasA (cce\_1751) interacts with rpaB instead of rpaA. Our analysis also gives a list of probable sensors that interact with rpaA, among which the most probable ones are cce\_0888 and cce\_2546. Notably, both of these have a high sequence similarity (32.083 percent identity, 3.45e-27 e-value for cce\_0888 and 32.308 percent identity, 2.33e-31 e-value for cce\_2546) with sasA of *Synechococcus*.

## cce\_1751/sasA

Cce\_1751 has the highest sequence similarity (43.716 percent identity, 1.26e-106 e-value) with the sasA kinase of *Synechococcus* among all genes in *Cyanothece*. Unlike *Synechococcus*, cce\_1751 is most closely associated with rpaB instead of rpaA. This prediction was consistent in both the datasets.

## cce\_4002/rpaB

From this analysis, it is clear that rpaB is a key regulator of the circadian output pathway along with rpaA because of its association with multiple clock genes, the sasA-like sensor cce\_1751, and multiple regulators. It has also been shown experimentally by Hanaoka et. al. (7) that RpaB binds to the kaiBC promoter and is a part of the circadian output pathway in *Synechococcus*.

## cce\_4751/cikA

Among all the two-component hybrid sensor and regulators evaluated, cce\_4751 was found to interact with a high mutual information score with multiple clock genes. Interestingly, this bears a high sequence similarity with the cikA gene of *Synechococcus* (43.06 percent identity, 4.86E-148 e-value).

## cce\_0888/nblS

Cce\_0888 is yet another interesting gene that is annotated as a two-component sensor histidine kinase and interacts with a high mutual information score with multiple clock genes and the regulator rpaA.

A Spearman Correlation based analysis of the genes of interest listed above along with the circadian clock genes is presented in Appendix C. This was done to further classify the mutual information-determined relationship into a positive (i.e., correlated gene expressions) or negative interaction (i.e., anti-correlated gene expressions).

# Conclusion

From the above analysis we conclude that:

* cce\_1983/aphA is a putative photoreceptor that interacts with multiple genes in the *Cyanothece* clock.
* The importance of cce\_0678 as a central response regulator is further highlighted in this study.
* rpaA and rpaB are equally important in the *Cyanothece* circadian pathway.
* Unlike *Synechococcus*, cce\_1751/sasA may not be the kinase that interacts with rpaA in *Cyanothece,* butinstead regulates rpaB.
* The multiple Kai copies present in *Cyanothece* might serve additional roles alongside maintaining robustness. These copies might be operating simultaneously to regulate different transcription factors employed for controlling different cellular processes. Our analysis indicates that both the *Cyanothece* clock module and the sensor-regulator pairs that interact with it do so in more complex manners than seen in *Synechococcus*. A majority of interactions seen in *Synechococcus* are conserved in *Cyanothece*, but a number of unique interactions are also predicted.

# Appendix A

The following table lists the E-values obtained by running a BLAST for a subset of the circadian clock module components of *Synechococcus* and maps them to *Cyanothece* genes.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***Synechococcus* gene** | ***Cyanothece* gene** | **Percent Identity** | **E-Value** | ***Cyanothece* homolog description** |
| CikA | cce\_4751 | 43.058 | 4.86e-148 | two-component hybrid sensor and regulator |
|  | cce\_4289 | 31.579 | 3.56e-68 | two-component hybrid sensor and regulator |
|  | cce\_1138 | 30.866 | 5.99e-60 | two-component hybrid sensor and regulator |
|  | cce\_0164 | 37.621 | 5.35e-59 | two-component sensor histidine kinase |
|  | cce\_0220 | 39.073 | 1.66e-53 | two-component sensor histidine kinase |
|  | cce\_2232 | 31.707 | 6.55e-42 | two-component sensor histidine kinase |
|  | cce\_1185 | 30.962 | 4.94e-45 | two-component hybrid sensor and regulator |
| SasA | cce\_1751 | 43.716 | 1.25e-106 | adaptive-response sensory histidine kinase |
|  | cce\_2546 | 32.308 | 2.33e-31 | two-component sensor histidine kinase |
|  | cce\_0888 | 32.083 | 3.45e-27 | two-component sensor histidine kinase |

# Appendix B

## Stockel Dataset

### cce\_0678

|  |  |  |  |
| --- | --- | --- | --- |
| **Gene Name** | **Functional Annotation** | **Common Name** | **Mutual Information** |
| cce\_1983 | probable phytochrome A, two-component sensor protein | aphA | 0.5079 |
| cce\_0220 | two-component sensor histidine kinase | cce\_0220 | 0.3718 |
| cce\_1535 | two-component sensor histidine kinase | cce\_1535 | 0.3587 |
| cce\_0969 | two-component sensor histidine kinase | cce\_0969 | 0.3397 |
| cce\_3894 | two-component sensor histidine kinase | cce\_3894 | 0.3321 |
| cce\_1878 | two-component sensor histidine kinase | cce\_1878 | 0.2729 |
| cce\_0297 | two-component sensor histidine kinase | cce\_0297 | 0.2640 |
| cce\_0016 | two-component sensor histidine kinase | cce\_0016 | 0.2529 |
| cce\_0164 | two-component sensor histidine kinase | cce\_0164 | 0.2502 |
| cce\_4006 | two-component sensor histidine kinase | cce\_4006 | 0.1980 |
| cce\_4204 | two-component sensor histidine kinase | cce\_4204 | 0.1804 |
| cce\_1280 | two-component sensor histidine kinase | cce\_1280 | 0.1362 |
| cce\_2232 | two-component sensor histidine kinase | cce\_2232 | 0.1232 |
| cce\_0257 | two-component sensor histidine kinase | cce\_0257 | 0.1229 |
| cce\_1519 | two-component sensor histidine kinase | cce\_1519 | 0.1170 |
| cce\_2366 | two-component sensor histidine kinase | cce\_2366 | 0.1035 |
| cce\_2546 | probable two-component sensor histidine kinase | cce\_2546 | 0.0952 |
| cce\_0888 | two-component sensor histidine kinase | nblS | 0.0839 |
| cce\_1467 | two-component sensor histidine kinase | cce\_1467 | 0.0726 |
| cce\_4097 | two-component sensor serine/threonine kinase | cce\_4097 | 0.0432 |
| cce\_1848 | probable two-component sensor histidine kinase | cce\_1848 | 0.0154 |

### rpaA

|  |  |  |  |
| --- | --- | --- | --- |
| **Gene Name** | **Functional Annotation** | **Common Name** | **Mutual Information** |
| cce\_0164 | two-component sensor histidine kinase | cce\_0164 | 0.6907 |
| cce\_1983 | probable phytochrome A, two-component sensor protein | aphA | 0.6170 |
| cce\_0888 | two-component sensor histidine kinase | nblS | 0.5268 |
| cce\_3379 | two-component sensor histidine kinase | cce\_3379 | 0.5044 |
| cce\_4426 | two-component sensor histidine kinase | cce\_4426 | 0.4777 |
| cce\_0297 | two-component sensor histidine kinase | cce\_0297 | 0.4613 |
| cce\_4097 | two-component sensor serine/threonine kinase | cce\_4097 | 0.4198 |
| cce\_2546 | probable two-component sensor histidine kinase | cce\_2546 | 0.3720 |
| cce\_0969 | two-component sensor histidine kinase | cce\_0969 | 0.2970 |
| cce\_2366 | two-component sensor histidine kinase | cce\_2366 | 0.2424 |
| cce\_3327 | two-component sensor histidine kinase | cce\_3327 | 0.2333 |
| cce\_3520 | two-component sensor histidine kinase | cce\_3520 | 0.2097 |
| cce\_0016 | two-component sensor histidine kinase | cce\_0016 | 0.1793 |
| cce\_0257 | two-component sensor histidine kinase | cce\_0257 | 0.1628 |
| cce\_1737 | two-component sensor histidine kinase | cce\_1737 | 0.1069 |
| cce\_0220 | two-component sensor histidine kinase | cce\_0220 | 0.0678 |
| cce\_1280 | two-component sensor histidine kinase | cce\_1280 | 0.0444 |
| cce\_4006 | two-component sensor histidine kinase | cce\_4006 | 0.0340 |
| cce\_1848 | probable two-component sensor histidine kinase | cce\_1848 | 0.0300 |
| cce\_1519 | two-component sensor histidine kinase | cce\_1519 | 0.0287 |
| cce\_4204 | two-component sensor histidine kinase | cce\_4204 | 0.0168 |
| cce\_1878 | two-component sensor histidine kinase | cce\_1878 | 0.0149 |
| cce\_3894 | two-component sensor histidine kinase | cce\_3894 | 0.0079 |

### sasA

|  |  |  |  |
| --- | --- | --- | --- |
| **Gene Name** | **Functional Annotation** | **Common Name** | **Mutual Information** |
| cce\_4002 | two-component response regulator | rpaB | 0.6551 |
| cce\_1808 | probable two-component system response regulator, OmpR subfamily | cce\_1808 | 0.5836 |
| cce\_4714 | two-component response regulator, NarL subfamily | cce\_4714 | 0.4452 |
| cce\_3895 | two-component response regulator | cce\_3895 | 0.4399 |
| cce\_0289 | two-component response regulator | cce\_0289 | 0.4093 |
| cce\_0678 | two-component response regulator | cce\_0678 | 0.3878 |
| cce\_1952 | two-component response regulator receiver protein | cce\_1952 | 0.3726 |
| cce\_3960 | two-component response regulator | cce\_3960 | 0.3599 |
| cce\_4416 | two-component response regulator | cce\_4416 | 0.3551 |
| cce\_3557 | two-component response regulator | cce\_3557 | 0.3482 |
| cce\_1135 | two-component response regulator | cce\_1135 | 0.3391 |
| cce\_3938 | two-component response regulator | cce\_3938 | 0.3266 |
| cce\_4183 | two-component response regulator | cce\_4183 | 0.2392 |
| cce\_2507 | two-component response regulator | cce\_2507 | 0.2147 |
| cce\_0298 | two-component response regulator | rpaA | 0.2135 |
| cce\_0970 | two-component transcription regulator | cce\_0970 | 0.1911 |
| cce\_3559 | two-component response regulator | cce\_3559 | 0.1704 |
| cce\_3714 | putative two-component system response regulator | cce\_3714 | 0.1217 |
| cce\_1520 | two-component response regulator | cce\_1520 | 0.1105 |
| cce\_3378 | two-component response regulator | cce\_3378 | 0.1014 |
| cce\_2050 | two-component system, regulatory protein | cce\_2050 | 0.0715 |
| cce\_4173 | two-component response regulator | cce\_4173 | 0.0670 |
| cce\_0712 | two-component response regulator | cce\_0712 | 0.0439 |
| cce\_3890 | probable two-component response regulator | cce\_3890 | 0.0340 |
| cce\_2365 | two-component response regulator | cce\_2365 | 0.0215 |
| cce\_0754 | two-component response regulator | cce\_0754 | 0.0093 |

### rpaB

|  |  |  |  |
| --- | --- | --- | --- |
| **Gene Name** | **Functional Annotation** | **Common Name** | **Mutual Information** |
| cce\_1751 | adaptive-response sensory histidine kinase | sasA | 0.6829 |
| cce\_1983 | probable phytochrome A, two-component sensor protein | aphA | 0.6288 |
| cce\_2366 | two-component sensor histidine kinase | cce\_2366 | 0.5451 |
| cce\_1982 | CheY-like two-component hybrid sensor and regulator | cce\_1982 | 0.4932 |
| cce\_1878 | two-component sensor histidine kinase | cce\_1878 | 0.4069 |
| cce\_3723 | two-component hybrid sensor and regulator | cce\_3723 | 0.3376 |
| cce\_0716 | two-component hybrid sensor and regulator | cce\_0716 | 0.2786 |
| cce\_0297 | two-component sensor histidine kinase | cce\_0297 | 0.2604 |
| cce\_2546 | probable two-component sensor histidine kinase | cce\_2546 | 0.2527 |
| cce\_1519 | two-component sensor histidine kinase | cce\_1519 | 0.2381 |
| cce\_1138 | two-component hybrid sensor and regulator | cce\_1138 | 0.2276 |
| cce\_0164 | two-component sensor histidine kinase | cce\_0164 | 0.2173 |
| cce\_4289 | two-component hybrid sensor and regulator | cce\_4289 | 0.1634 |
| cce\_2375 | two-component hybrid sensor and regulator | cce\_2375 | 0.1448 |
| cce\_4195 | CheY-like two-component hybrid sensor and regulator | cce\_4195 | 0.1388 |
| cce\_0888 | two-component sensor histidine kinase | nblS | 0.1369 |
| cce\_1737 | two-component sensor histidine kinase | cce\_1737 | 0.1361 |
| cce\_4006 | two-component sensor histidine kinase | cce\_4006 | 0.1297 |
| cce\_3937 | two-component hybrid sensor and regulator | cce\_3937 | 0.1244 |
| cce\_3894 | two-component sensor histidine kinase | cce\_3894 | 0.1205 |
| cce\_0220 | two-component sensor histidine kinase | cce\_0220 | 0.1100 |
| cce\_2505 | two-component hybrid sensor and regulator | cce\_2505 | 0.0988 |
| cce\_0656 | two-component hybrid sensor and regulator | cce\_0656 | 0.0984 |
| cce\_0016 | two-component sensor histidine kinase | cce\_0016 | 0.0934 |
| cce\_1777 | probable two-component hybrid sensor and regulator | cce\_1777 | 0.0888 |
| cce\_3722 | two-component hybrid sensor and regulator | cce\_3722 | 0.0823 |
| cce\_4204 | two-component sensor histidine kinase | cce\_4204 | 0.0767 |
| cce\_1413 | probable sensory transduction histidine kinase | cce\_1413 | 0.0729 |
| cce\_3379 | two-component sensor histidine kinase | cce\_3379 | 0.0654 |
| cce\_3558 | two-component hybrid sensor and regulator | cce\_3558 | 0.0629 |
| cce\_1280 | two-component sensor histidine kinase | cce\_1280 | 0.0304 |
| cce\_1694 | two component sensor histidine kinase | cce\_1694 | 0.0069 |
| cce\_1467 | two-component sensor histidine kinase | cce\_1467 | 0.0064 |
| cce\_2508 | two-component hybrid sensor and regulator | cce\_2508 | 0.0037 |
| cce\_2624 | CheA-like two-component hybrid sensor and regulator | cce\_2624 | 0.0027 |

## Toepel Dataset

### cce\_0678

|  |  |  |  |
| --- | --- | --- | --- |
| **Gene Name** | **Functional Annotation** | **Common Name** | **Mutual Information** |
| cce\_0164 | two-component sensor histidine kinase | cce\_0164 | 0.3980 |
| cce\_4426 | two-component sensor histidine kinase | cce\_4426 | 0.2421 |
| cce\_1280 | two-component sensor histidine kinase | cce\_1280 | 0.2249 |
| cce\_2232 | two-component sensor histidine kinase | cce\_2232 | 0.1923 |
| cce\_0220 | two-component sensor histidine kinase | cce\_0220 | 0.1889 |
| cce\_1983 | probable phytochrome A, two-component sensor protein | aphA | 0.1718 |
| cce\_2546 | probable two-component sensor histidine kinase | cce\_2546 | 0.1640 |
| cce\_1878 | two-component sensor histidine kinase | cce\_1878 | 0.1289 |
| cce\_1535 | two-component sensor histidine kinase | cce\_1535 | 0.1258 |
| cce\_0888 | two-component sensor histidine kinase | nblS | 0.1209 |
| cce\_0257 | two-component sensor histidine kinase | cce\_0257 | 0.1024 |
| cce\_3894 | two-component sensor histidine kinase | cce\_3894 | 0.1007 |
| cce\_0969 | two-component sensor histidine kinase | cce\_0969 | 0.0971 |
| cce\_1467 | two-component sensor histidine kinase | cce\_1467 | 0.0746 |
| cce\_0297 | two-component sensor histidine kinase | cce\_0297 | 0.0632 |
| cce\_3327 | two-component sensor histidine kinase | cce\_3327 | 0.0586 |
| cce\_3520 | two-component sensor histidine kinase | cce\_3520 | 0.0393 |
| cce\_4097 | two-component sensor serine/threonine kinase | cce\_4097 | 0.0295 |
| cce\_1848 | probable two-component sensor histidine kinase | cce\_1848 | 0.0099 |

### rpaA

|  |  |  |  |
| --- | --- | --- | --- |
| **Gene Name** | **Functional Annotation** | **Common Name** | **Mutual Information** |
| cce\_4097 | two-component sensor serine/threonine kinase | cce\_4097 | 0.5239 |
| cce\_0220 | two-component sensor histidine kinase | cce\_0220 | 0.4440 |
| cce\_1280 | two-component sensor histidine kinase | cce\_1280 | 0.4273 |
| cce\_0297 | two-component sensor histidine kinase | cce\_0297 | 0.4259 |
| cce\_3379 | two-component sensor histidine kinase | cce\_3379 | 0.3912 |
| cce\_4426 | two-component sensor histidine kinase | cce\_4426 | 0.3457 |
| cce\_3327 | two-component sensor histidine kinase | cce\_3327 | 0.2290 |
| cce\_1519 | two-component sensor histidine kinase | cce\_1519 | 0.1655 |
| cce\_2546 | probable two-component sensor histidine kinase | cce\_2546 | 0.0865 |
| cce\_4204 | two-component sensor histidine kinase | cce\_4204 | 0.0853 |
| cce\_4006 | two-component sensor histidine kinase | cce\_4006 | 0.0695 |
| cce\_3520 | two-component sensor histidine kinase | cce\_3520 | 0.0564 |
| cce\_1535 | two-component sensor histidine kinase | cce\_1535 | 0.0551 |
| cce\_2232 | two-component sensor histidine kinase | cce\_2232 | 0.0402 |
| cce\_1737 | two-component sensor histidine kinase | cce\_1737 | 0.0374 |
| cce\_0969 | two-component sensor histidine kinase | cce\_0969 | 0.0245 |
| cce\_0164 | two-component sensor histidine kinase | cce\_0164 | 0.0112 |
| cce\_1878 | two-component sensor histidine kinase | cce\_1878 | 0.0057 |

### sasA

|  |  |  |  |
| --- | --- | --- | --- |
| **Gene Name** | **Functional Annotation** | **Common Name** | **Mutual Information** |
| cce\_1952 | two-component response regulator receiver protein | cce\_1952 | 0.4666 |
| cce\_4002 | two-component response regulator | rpaB | 0.3694 |
| cce\_0712 | two-component response regulator | cce\_0712 | 0.3392 |
| cce\_4578 | two-component response regulator | cce\_4578 | 0.2911 |
| cce\_4714 | two-component response regulator, NarL subfamily | cce\_4714 | 0.2750 |
| cce\_2376 | two-component response regulator | cce\_2376 | 0.1687 |
| cce\_1695 | two-component response regulator | cce\_1695 | 0.1547 |
| cce\_2365 | two-component response regulator | cce\_2365 | 0.1271 |
| cce\_3895 | two-component response regulator | cce\_3895 | 0.0970 |
| cce\_3714 | putative two-component system response regulator | cce\_3714 | 0.0948 |
| cce\_3557 | two-component response regulator | cce\_3557 | 0.0897 |
| cce\_0678 | two-component response regulator | cce\_0678 | 0.0843 |
| cce\_0970 | two-component transcription regulator | cce\_0970 | 0.0813 |
| cce\_3190 | two-component response regulator | cce\_3190 | 0.0674 |
| cce\_0446 | two-component response regulator | cce\_0446 | 0.0361 |

### rpaB

|  |  |  |  |
| --- | --- | --- | --- |
| **Gene Name** | **Functional Annotation** | **Common Name** | **Mutual Information** |
| cce\_1878 | two-component sensor histidine kinase | cce\_1878 | 0.4266 |
| cce\_1751 | adaptive-response sensory histidine kinase | sasA | 0.3694 |
| cce\_2505 | two-component hybrid sensor and regulator | cce\_2505 | 0.3681 |
| cce\_3723 | two-component hybrid sensor and regulator | cce\_3723 | 0.3446 |
| cce\_0257 | two-component sensor histidine kinase | cce\_0257 | 0.3398 |
| cce\_0164 | two-component sensor histidine kinase | cce\_0164 | 0.2424 |
| cce\_1777 | probable two-component hybrid sensor and regulator | cce\_1777 | 0.2044 |
| cce\_3558 | two-component hybrid sensor and regulator | cce\_3558 | 0.1913 |
| cce\_0016 | two-component sensor histidine kinase | cce\_0016 | 0.1173 |
| cce\_2375 | two-component hybrid sensor and regulator | cce\_2375 | 0.1080 |
| cce\_2232 | two-component sensor histidine kinase | cce\_2232 | 0.1060 |
| cce\_0656 | two-component hybrid sensor and regulator | cce\_0656 | 0.0916 |
| cce\_0220 | two-component sensor histidine kinase | cce\_0220 | 0.0914 |
| cce\_1467 | two-component sensor histidine kinase | cce\_1467 | 0.0861 |
| cce\_1982 | CheY-like two-component hybrid sensor and regulator | cce\_1982 | 0.0785 |
| cce\_4751 | two-component hybrid sensor and regulator | cce\_4751 | 0.0780 |
| cce\_1983 | probable phytochrome A, two-component sensor protein | aphA | 0.0693 |
| cce\_4204 | two-component sensor histidine kinase | cce\_4204 | 0.0615 |
| cce\_4405 | two-component hybrid sensor and regulator | cce\_4405 | 0.0556 |
| cce\_1185 | two-component hybrid sensor and regulator | cce\_1185 | 0.0480 |
| cce\_1556 | putative sensor protein | cce\_1556 | 0.0467 |
| cce\_1848 | probable two-component sensor histidine kinase | cce\_1848 | 0.0432 |
| cce\_0297 | two-component sensor histidine kinase | cce\_0297 | 0.0286 |
| cce\_2624 | CheA-like two-component hybrid sensor and regulator | cce\_2624 | 0.0267 |
| cce\_3520 | two-component sensor histidine kinase | cce\_3520 | 0.0242 |
| cce\_4426 | two-component sensor histidine kinase | cce\_4426 | 0.0185 |

# Appendix C

## Stockel Dataset

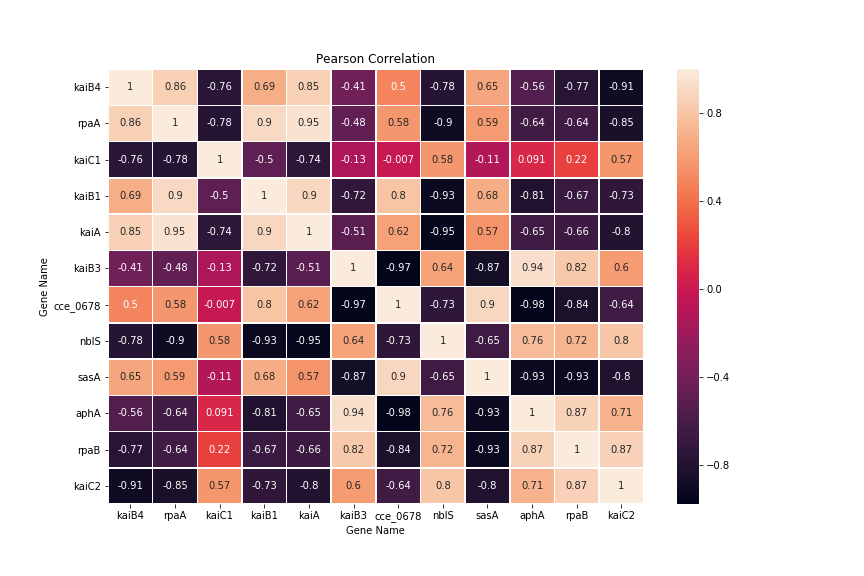


Figure 9: A Spearman Correlation based study of the genes of interest using the expression profiles of these genes from the Stockel Dataset. Each element of a matrix represents the correlation value between the genes of that particular row and column.

## Toepel Dataset

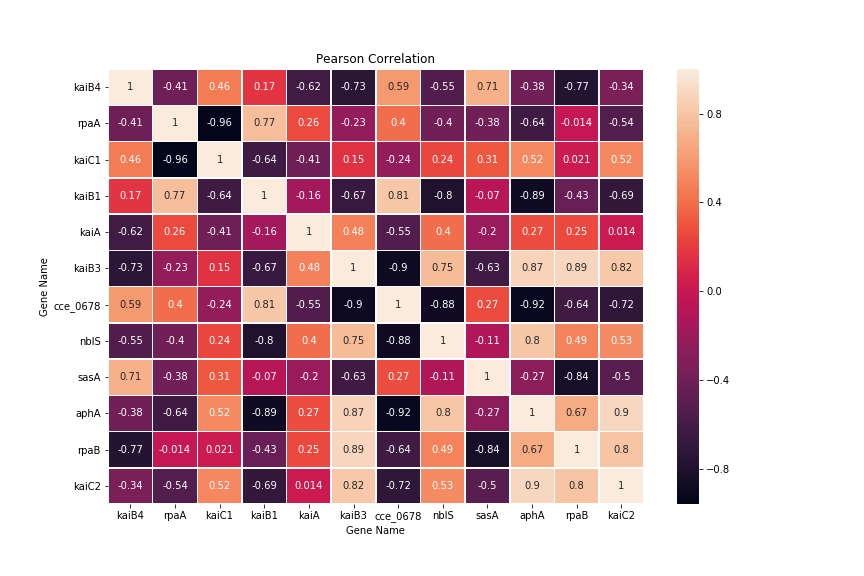


Figure 10: A Spearman Correlation based study of the genes of interest using the expression profiles of these genes from the Toepel Dataset. Each element of a matrix represents the correlation value between the genes of that particular row and column.

# References

1. *Circadian Rhythms in Cyanobacteria.* **Cohen, Susan.** 4, Microbiology and Molecular Biology Reviews, Vol. 79, pp. 373–385.

2. *A model of the circadian clock in the cyanobacterium Cyanothece sp. ATCC 51142.* **Vinh, Nguyen Xuan, et al.** 2, 2013, BMC Bioinformatics, Vol. 14, pp. 1-9.

3. *Mutual information and k-nearest neighbors approximator for time series prediction.* **Sorjamaa, Antti, Hao, Jin and Lendasse, Amaury.** 2005, Lecture Notes in Computer Science, pp. 553-558.

4. *Global transcriptomic analysis of Cyanothece 51142 reveals robust diurnal oscillation of central metabolic processes.* **Stöckel, Jana, et al.** 16, 2008, Proceedings of the National Academy of Sciences of the United States of America, Vol. 105, pp. 6156-6161.

5. *Differential Transcriptional Analysis of the Cyanobacterium Cyanothece sp. Strain ATCC 51142 during Light-Dark and Continuous-Light Growth.* **Toepel, Jörg, et al.** 11, 2008, Journal of Bacteriology, Vol. 190, pp. 3904-3913.

6. *The genome of Cyanothece 51142, a unicellular diazotrophic cyanobacterium important in the marine nitrogen cycle.* **Welsh, Eric A., et al.** 39, 2008, Proceedings of the National Academy of Sciences of the United States of America, Vol. 105, pp. 15094-15099.

7. **Hanaoka, Mitsumasa and Tanaka, Kan.** *RpaB, another response regulator operating circadian clock-dependent transcriptional regulation in Synechococcus elongatus PCC 7942.* s.l. : Journal of Biological Chemistry, 2012.