Understanding *Cyanothece* 51142 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 external environmental cues 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 much 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. We will use Mutual Information as a correlation metric because of its ability to catch non-linear interaction between variables as opposed to Pearson correlation which only captures linear correlation. In addition to finding out the sensors and regulators that are highly correlated with the clock genes, we can also infer whether a sensor-regulator pair belongsto the same two-component system. Since they are a part of the same unit, their expression profiles must also be very highly correlated. This way, we can gain more insight into the clock module of *Cyanothece* by using Mutual Information.

# Mutual Information

If we consider two random variables X and Y, the Mutual Information between them would be

where H(.) computes the Shannon Entropy. For continuous variables, the idea is to estimate H(.) from the average distance to the k nearest neighbors (3). MI is 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 i containing k nearest neighbour.

# MicroArray Dataset

The microarray expression datasets are 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 later used to verify the results.

# Analysis

## Stockel Dataset

The expression profile of the genes in the Stockel Dataset were first filtered to include only those genes that are annotated as circadian clock protein or two-component sensors and regulators. The gene to annotation mapping was obtained from the genomic database submitted by Welsh et. al. (6). Using the expression profiles of those selected genes, the mutual information between the individual clock genes and the sensors and regulators were obtained. To visualize the interactions, correlation matrices were created as shown below. The X-axis contains the Kai clock genes and the y-axis contains the functionally annotated sensors and regulators in *Cyanothece*. The value corresponding to the x\_i th column and the y\_ith row represents the mutual information score between the th Kai Gene and the th sensor/regulator. The interaction of the clock proteins with the two-component sensors, regulators and the two-component hybrid sensor & regulator is geiven in Figure 1, Figure 2 and Figure 3 respectively.

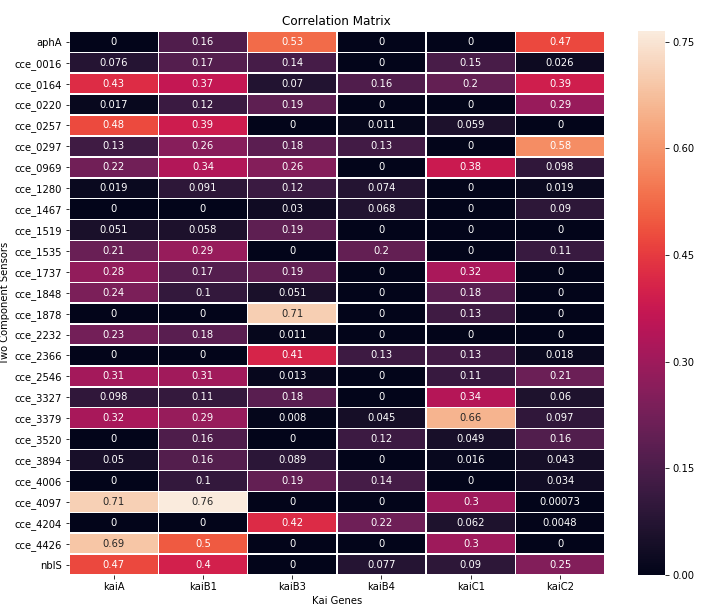


Figure 1

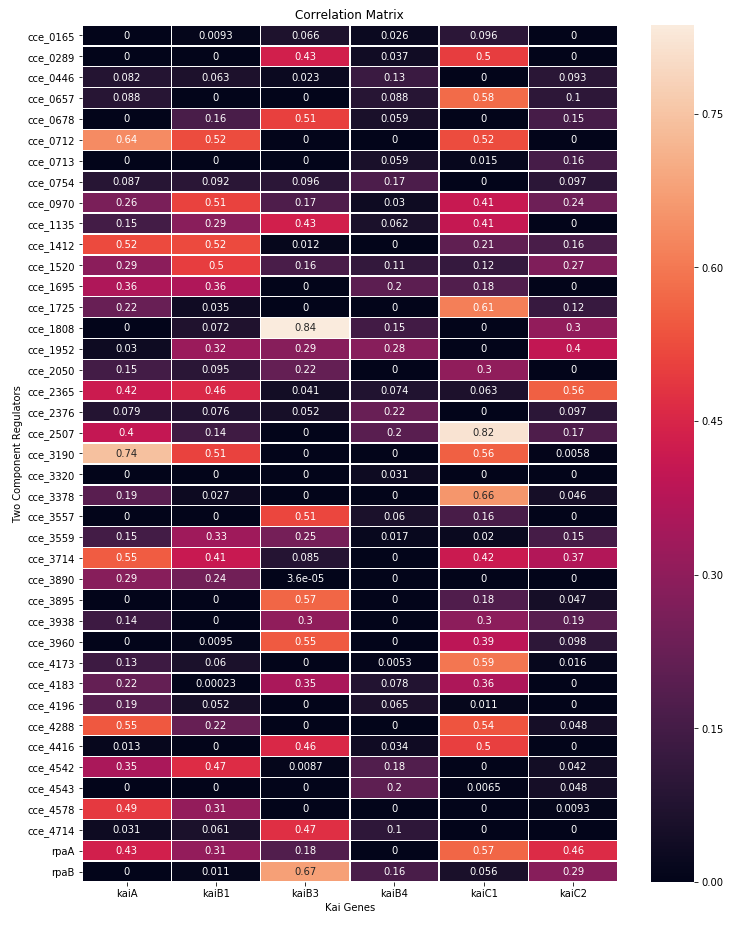


Figure 2

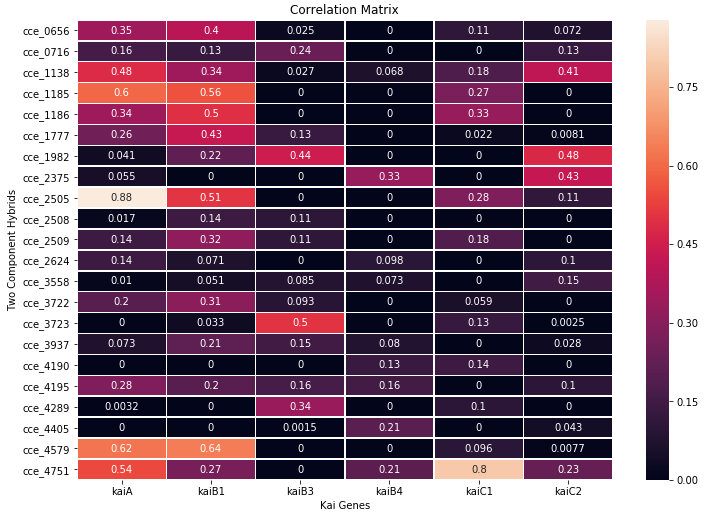


Figure 3

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

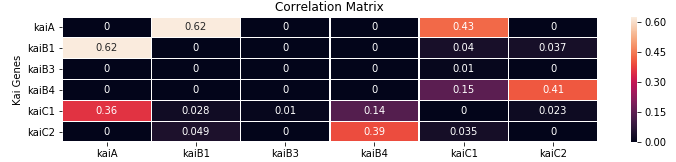


Figure 4

## 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.

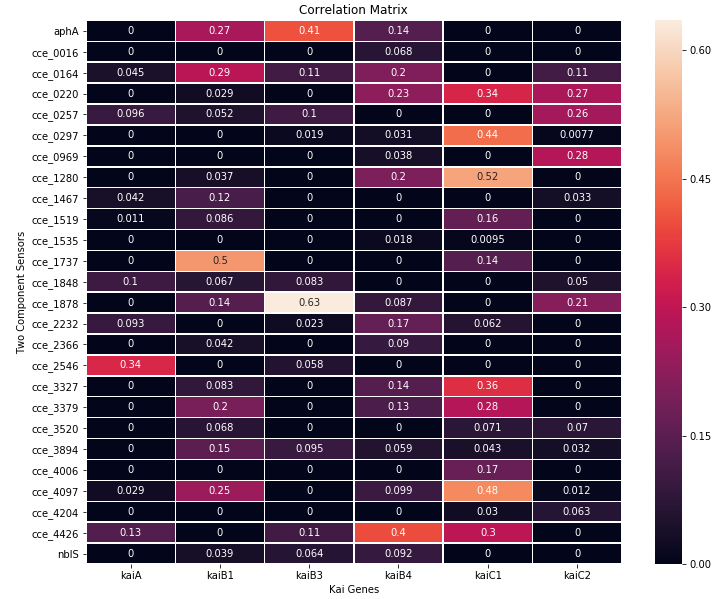


Figure 5

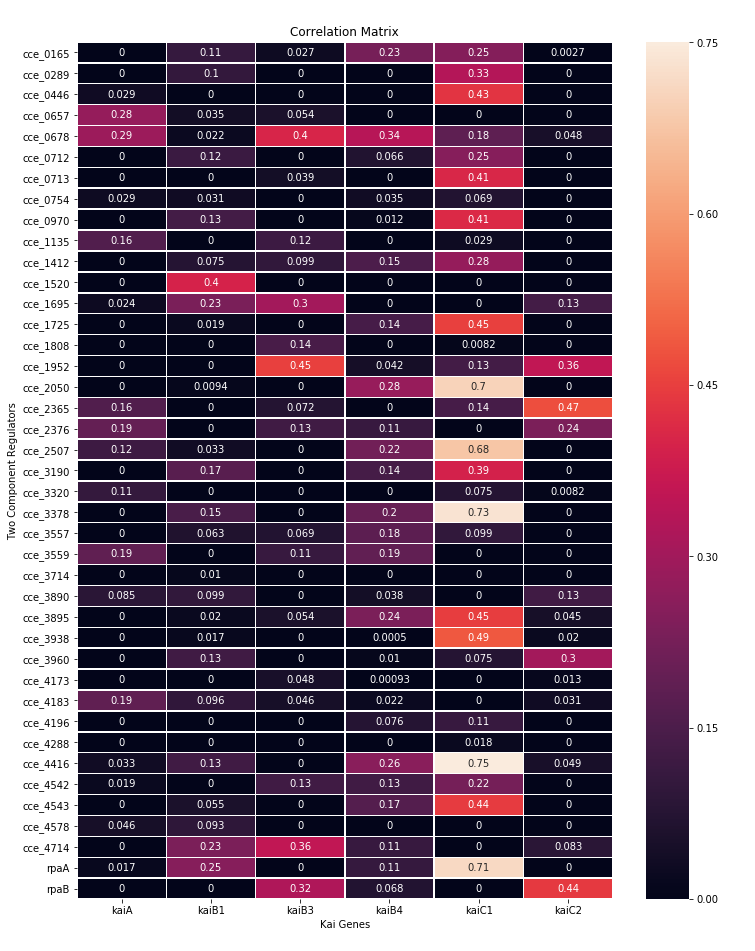


Figure 6

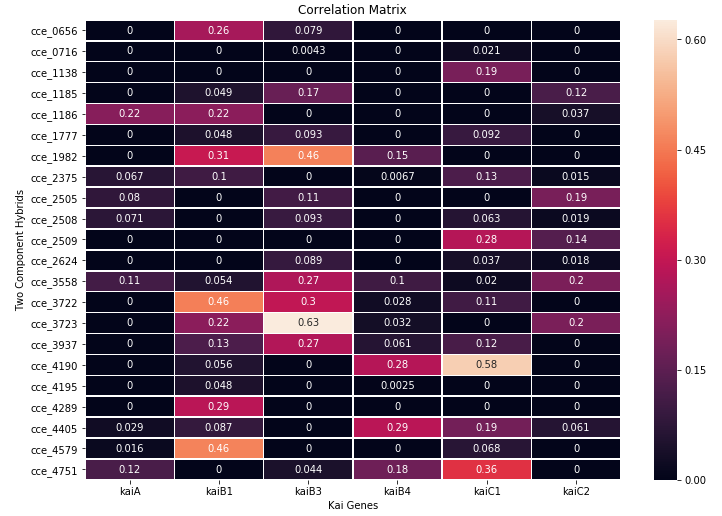


Figure 7

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

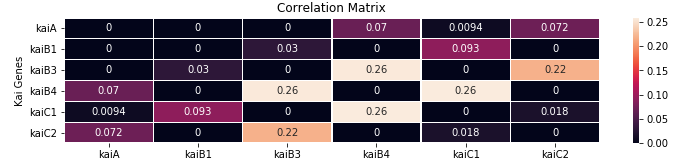


Figure 8

# Finding the most active set of KaiABC combination

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

## Stockel Dataset

The table below presents the KaiABC combinations and the number of common sensors, regulators and hybrid sensors & regulators that they are interact with. The table is arranged in descending order of the total number of common sensors,regulators and hybrids that the given KaiABC combination interacts with.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Kai Combination | No. of sensors | No. of regulators | No. of hybrids | Total |
| kaiAB1C1 | 13 | 19 | 19 | 51 |
| kaiAB1C2 | 13 | 19 | 19 | 51 |
| kaiAB3C2 | 10 | 13 | 13 | 36 |
| kaiAB3C1 | 9 | 12 | 12 | 33 |
| kaiAB4C2 | 6 | 12 | 12 | 30 |
| kaiAB4C1 | 4 | 12 | 12 | 28 |

## Toepel Dataset

A similar table as above is presented below using the Toepel Dataset.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Kai Combination | No. of sensors | No. of regulators | No. of hybrids | Total |
| kaiAB1C2 | 5 | 5 | 5 | 15 |
| kaiAB4C1 | 3 | 6 | 6 | 15 |
| kaiAB3C2 | 3 | 5 | 5 | 13 |
| kaiAB1C1 | 2 | 5 | 5 | 12 |
| kaiAB4C2 | 2 | 5 | 5 | 12 |
| kaiAB3C1 | 2 | 4 | 4 | 10 |

# Case Study of some genes of interest

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

## cce\_1983/aphA

Cce\_1983 has been shown to have a very high mutual information with multiple clock genes in both studies using Stockel or 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 has previously been reported to share a high correlation with the RubisCo genes. In this study it is seen to share a high mutual information score with a number of important genes like cce\_1983, cce\_0220, cce\_0164, cce\_2232 which have all been annotated as sensors. 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. The E-values of the BLAST search results are all presented in Appendix A.

## cce\_0298/rpaA

Cce\_0298/rpaA has been shown to be a master regulator in *Synechococcus*. In this study as well, it is seen 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 of *Synechococcus*, does not have any correlation with rpaA. On the other hand, cce\_1751 has a very high mutual information score with rpaB or cce\_4002. Therefore, sasA or cce\_1751 may not be the sensor that interacts with rpaA. Our analysis gives a list of probable sensors that interact with rpaA, among which the most probable ones are cce\_0888 and cce\_2546 both of which also have a high sequence similarity with sasA of *Synechococcus*.

## cce\_1751/sasA

Cce\_1751 has the highest sequence similarity with the sasA kinase of *Synechococcus* among all genes in *Cyanothece*. Unlike *Synechococcus* however, cce\_1751 is most closely associated with rpaB instead of rpaA as evident from the analysis of both the datasets.

## cce\_4002/rpaB

From this analysis, it is clear that rpaB is also a key regulator of the circadian output pathway along with rpaA because of its association with multiple clock genes, 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 the two-component hybrid sensor & regulators, cce\_4751 stands out because of its high correlation with multiple clock genes and high sequence similarity with the cikA gene of *Synechococcus*.

## cce\_0888/nblS

Cce\_0888 is another interesting gene that is anotated as a two-component sensor histidine kinase and correlates highly with multiple clock genes and the regulator rpaA as seen in the analysis.

A Spearman Correlation based analysis of the genes of interest listed above along with the circadian clock genes is presented in Appendix C. This is done to find the positive or negative nature of interaction between them.

# Conclusion

From the above analysis we can conclude that:

* cce\_1983/aphA might be a photoreceptor that interacts with the clock.
* cce\_0678's importance as a regulator is further highlighted in this study.
* rpaA and rpaB are equally important in the *Cyanothece* circadian output pathway.
* cce\_1751/sasA may not be the kinase that interacts with rpaA in *Cyanothece*. On the other hand, it may regulate rpaB.
* The Kai copies may be present not just to maintain robustness. There may be two separate clocks that operate simultaneosly and regulate similar transcription factors for different processes. This can be the reason why *Cyanothece* has multiple copies of not only the clock genes but also the kinases and regulators that are supposed to interact with the clock. Maybe that's how they separate two conflicting processes, Photosynthesis and Nitrogen Fixation.

# Appendix A

The following table lists the E-values obtained by running a BLAST search of some circadian clock module components of *Synechococcus* and maps them to *Cyanothece* genes. It is obtained from the work of Vinh et. al. (2).

|  |  |  |  |
| --- | --- | --- | --- |
| *Synechococcus* | *Cyanothece* | E-Value | *Cyanothece* homolog description |
| CikA | cce\_4751 | 1e-129 | two-component hybrid sensor and regulator |
|  | cce\_4289 | 7e-67 | two-component hybrid sensor and regulator |
|  | cce\_1138 | 2e-59 | two-component hybrid sensor and regulator |
|  | cce\_0164 | 1e-52 | two-component sensor histidine kinase |
|  | cce\_0220 | 4e-52 | two-component sensor histidine kinase |
|  | cce\_2232 | 3e-46 | two-component sensor histidine kinase |
|  | cce\_1185 | 7e-46 | two-component hybrid sensor and regulator |
| SasA | cce\_1751 | 9e-81 | adaptive-response sensory histidine kinase |
|  | cce\_2546 | 3e-27 | two-component sensor histidine kinase |
|  | cce\_0888 | 4e-25 | two-component sensor histidine kinase |