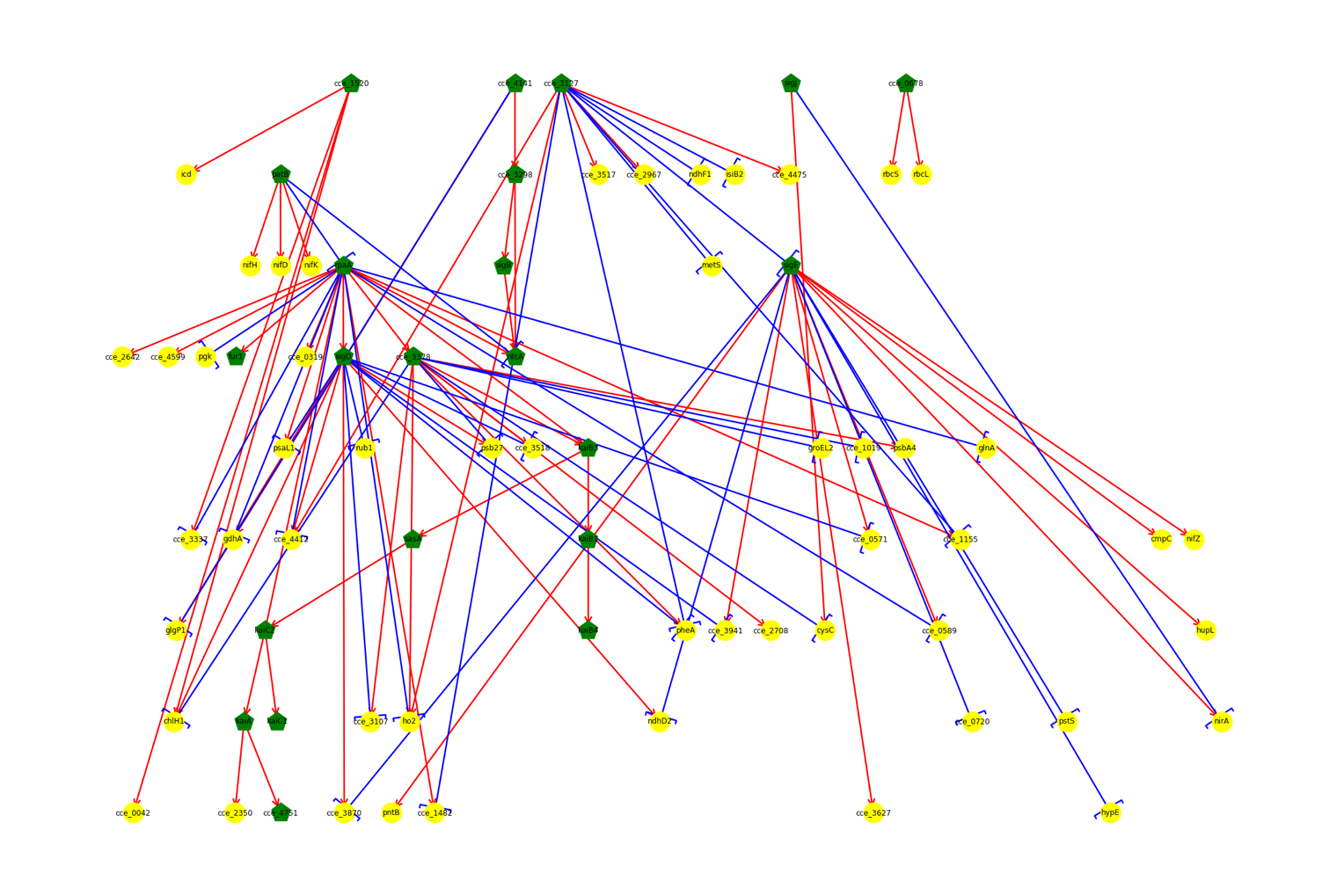
# Report on the Gene Regulatory Network of *Cyanothece* sp. 51142

The following network depicts the gene regulatory network of *Cyanothece* sp. 51142 constructed from literature sources. A complete list of the sources along with the algorithms used to infer the edges is provided in a later section of this report.



Figure

*To better visualize the network, an image file named pGRN.png has also been provided along with this report.*

## Network and it's description

The network has the following components:

1. The Transcription Factors indicated as the green pentagons.

2. The Target Genes indicated as the yellow ellipses.

3. A red arrow denotes an activating edge.

4. A blue arrow denotes an inhibiting edge.

A list of all the transcription factors which act as source nodes in the network topology along with their annotated functions is presented below. The gene names and their annotations are obtained from the sequenced genomic data provided by Welsh et. al. [1].

Table

|  |  |
| --- | --- |
| Transcription Factors | Annotated Functions |
| cce\_0678 | two-component response regulator |
| sigE | group 2 sigma-70 RNA polymerase sigma factor E |
| cce\_4141 | conserved hypothetical protein |
| kaiA | circadian clock protein |
| sasA | adaptive-response sensory histidine kinase |
| cce\_3127 | transcription regulator, Fur family |
| cce\_1520 | two-component response regulator |
| kaiB1 | circadian clock protein |
| patB | probable transcriptional regulator |
| sigD | group 2 sigma-70 RNA polymerase sigma factor D |
| cce\_3378 | two-component response regulator |
| sigB | group 2 sigma-70 RNA polymerase sigma factor B |
| rpaA | two-component response regulator |
| kaiC2 | circadian clock protein |
| sigJ | group 3 sigma-70 RNA polymerase sigma factor J |
| cce\_3298 | unknown |
| kaiB3 | circadian clock protein |

A list of all the target genes/transcription factors which solely act as target nodes in the network topology along with their annotated functions is presented below.

Table

|  |  |
| --- | --- |
| Target Genes | Annotated Functions |
| cysC | adenylylsulfate kinase |
| cce\_1482 | conserved hypothetical protein |
| rub1 | rubredoxin |
| nifK | nitrogenase molybdenum-iron protein beta chain |
| nifH | nitrogenase iron protein |
| pntB | pyridine nucleotide transhydrogenase beta subunit |
| ntcA | nitrogen-responsive regulatory protein |
| ho2 | heme oxygenase |
| chlH1 | magnesium chelatase, subunit H |
| groEL2 | chaperonin 2 |
| cce\_3517 | putative S-layer OprB family carbohydrate-selective porin |
| cce\_4751 | two-component hybrid sensor and regulator |
| nifZ | iron-sulfur cofactor synthesis protein |
| cce\_2967 | magnesium-protoporphyrin IX monomethyl ester aerobic oxidative cyclase |
| cce\_1155 | Peptidase M48, Ste24p |
| cce\_4599 | putative cytochrome oxidase assembly |
| pstS | phosphate ABC transporter, periplasmic phosphate-binding protein |
| metS | methionyl-tRNA synthetase |
| cce\_0042 | histidinol-phosphate phosphatase, HAD-superfamily hydrolase subfamily IIIA |
| fur1 | ferric uptake regulation protein |
| cce\_4475 | putative arsenical pump-driving ATPase |
| cce\_2350 | putative alpha-helical ferredoxin |
| cce\_2642 | circadian phase modifier CpmA-like protein |
| rbcL | ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit |
| hupL | [NiFe |
| icd | isocitrate dehydrogenase (NADP+) |
| psaL1 | photosystem I reaction centre subunit XI PsaL |
| cce\_3337 | putative Glycosyl transferase, group 1 |
| gdhA | glutamate dehydrogenase |
| kaiB4 | putative circadian clock protein |
| cmpC | bicarbontate transport system ATP-binding protein |
| cce\_0720 | putative peptidoglycan-binding peptidase, M23B family |
| isiB2 | flavodoxin |
| cce\_0319 | putative S-layer OprB family carbohydrate-selective porin |
| ndhD2 | NADH dehydrogenase subunit 4 |
| cce\_0571 | 2Fe-2S ferredoxin, putative nitrogen fixation related protein |
| rbcS | ribulose bisphosphate carboxylase small subunit |
| cce\_3107 | elongation factor EF-G fragment |
| psbA4 | photosystem II D1 protein |
| nifD | nitrogenase molybdenum-iron protein alpha chain |
| pheA | prephenate dehydratase |
| cce\_0589 | cytochrome c family protein |
| cce\_4412 | magnesium-protoporphyrin IX monomethyl ester aerobic oxidative cyclase |
| cce\_2708 | rfrA family pentapeptide repeat |
| cce\_1019 | glycosyl transferase, group 1 |
| cce\_3941 | alpha amylase |
| cce\_3870 | sulfatase |
| ndhF1 | NADH dehydrogenase subunit 5 |
| glnA | glutamine synthetase, glutamate--ammonia ligase |
| kaiC1 | circadian clock protein |
| cce\_3627 | DUF1257-containing protein |
| hypE | hydrogenase expression/formation protein |
| pgk | phosphoglycerate kinase |
| psb27 | photosystem II 11 kD protein |
| nirA | ferredoxin-nitrite reductase |
| cce\_3518 | putative S-layer OprB family carbohydrate-selective porin |
| glgP1 | glycogen phosphorylase |

## Edge Elucidation

The following table contains the list of edges in the network along with the weight on every edge and their literature source annotations. A weight greater than 0 denotes activation while a weight less than 0 denotes inhibition. The literature source annotations are provided to give a quick reference to the algorithm that is used in that particular literature to infer a relation between a transcription factor and its corresponding target gene. A mapping between the literature source and the respective algorithm used is given in a later table.

Table

|  |  |  |  |
| --- | --- | --- | --- |
| Transcription Factor | Target Gene | Weight | Literature Source |
| rpaA | kaiB3 | 1 | Vinh et. al. [2] |
| rpaA | cce\_2642 | 1 | Vinh et. al. [2] |
| rpaA | cce\_3378 | 1 | McDermott et. al. [3] |
| rpaA | fur1 | 1 | McDermott et. al. [3] |
| rpaA | sigD | 1 | McDermott et. al. [3] |
| rpaA | ntcA | 1 | McDermott et. al. [3] |
| rpaA | cce\_4599 | 0.29744 | Mueller et. al. [4] |
| rpaA | glnA | -0.1663 | Mueller et. al. [4] |
| rpaA | cce\_4412 | -0.65524 | Mueller et. al. [4] |
| rpaA | pgk | -0.31148 | Mueller et. al. [4] |
| rpaA | psaL1 | 0.41655 | Mueller et. al. [4] |
| rpaA | cce\_3337 | -0.07848 | Mueller et. al. [4] |
| rpaA | ho2 | -0.76869 | Mueller et. al. [4] |
| rpaA | cce\_1482 | 0.15361 | Mueller et. al. [4] |
| rpaA | cce\_1155 | 0.11268 | Mueller et. al. [4] |
| rpaA | gdhA | -0.19142 | Mueller et. al. [4] |
| rpaA | cce\_0589 | -0.40736 | Mueller et. al. [4] |
| rpaA | cce\_0319 | 0.15584 | Mueller et. al. [4] |
| sigE | cce\_3941 | 0.26304 | Mueller et. al. [4] |
| sigE | cce\_3870 | -0.16728 | Mueller et. al. [4] |
| sigE | cce\_3627 | 0.48201 | Mueller et. al. [4] |
| sigE | ndhD2 | -0.55367 | Mueller et. al. [4] |
| sigE | hypE | -0.14227 | Mueller et. al. [4] |
| sigE | nirA | 0.54975 | Mueller et. al. [4] |
| sigE | hupL | 0.317 | Mueller et. al. [4] |
| sigE | pstS | -0.38369 | Mueller et. al. [4] |
| sigE | pntB | 0.41242 | Mueller et. al. [4] |
| sigE | cce\_0720 | -0.23997 | Mueller et. al. [4] |
| sigE | cce\_0589 | 0.1214 | Mueller et. al. [4] |
| sigE | cce\_0571 | 0.61694 | Mueller et. al. [4] |
| sigE | nifZ | 0.20534 | Mueller et. al. [4] |
| sigE | cmpC | 0.58494 | Mueller et. al. [4] |
| cce\_0678 | rbcS | 1 | McDermott et. al. [3] |
| cce\_0678 | rbcL | 1 | McDermott et. al. [3] |
| kaiA | cce\_2350 | 1 | Vinh et. al. [2] |
| kaiA | cce\_4751 | 1 | Vinh et. al. [2] |
| cce\_3378 | pheA | 0.24196 | Mueller et. al. [4] |
| cce\_3378 | chlH1 | -0.52655 | Mueller et. al. [4] |
| cce\_3378 | cysC | -0.20478 | Mueller et. al. [4] |
| cce\_3378 | psb27 | -0.32931 | Mueller et. al. [4] |
| cce\_3378 | cce\_3518 | 0.45845 | Mueller et. al. [4] |
| cce\_3378 | psbA4 | 0.25356 | Mueller et. al. [4] |
| cce\_3378 | groEL2 | -0.35666 | Mueller et. al. [4] |
| cce\_3378 | cce\_3107 | 0.52238 | Mueller et. al. [4] |
| cce\_3378 | cce\_2708 | 0.18337 | Mueller et. al. [4] |
| cce\_3378 | ho2 | 0.43344 | Mueller et. al. [4] |
| cce\_3378 | cce\_1019 | -0.06623 | Mueller et. al. [4] |
| cce\_3378 | kaiB3 | 0.23836 | Mueller et. al. [4] |
| kaiB1 | kaiB4 | 1 | Vinh et. al. [2] |
| sasA | kaiC2 | 1 | Vinh et. al. [2] |
| sigJ | cysC | 0.14945 | Mueller et. al. [4] |
| sigJ | nirA | -0.067 | Mueller et. al. [4] |
| cce\_3298 | sigB | 1 | McDermott et. al. [3] |
| cce\_3298 | ntcA | 1 | McDermott et. al. [3] |
| cce\_1520 | chlH1 | 0.33167 | Mueller et. al. [4] |
| cce\_1520 | cce\_3337 | 0.28858 | Mueller et. al. [4] |
| cce\_1520 | icd | 0.27024 | Mueller et. al. [4] |
| cce\_1520 | cce\_0042 | 0.1062 | Mueller et. al. [4] |
| kaiB3 | kaiB1 | 1 | Vinh et. al. [2] |
| kaiB3 | sasA | 1 | Vinh et. al. [2] |
| cce\_4141 | cce\_3298 | 1 | McDermott et. al. [3] |
| cce\_4141 | glgP1 | -0.44205 | Mueller et. al. [4] |
| cce\_4141 | gdhA | 0.22246 | Mueller et. al. [4] |
| kaiC2 | kaiA | 1 | Vinh et. al. [2] |
| kaiC2 | kaiC1 | 1 | Vinh et. al. [2] |
| kaiC2 | rpaA | 1 | Vinh et. al. [2] |
| cce\_3127 | pheA | -0.34813 | Mueller et. al. [4] |
| cce\_3127 | cce\_4475 | 0.05345 | Mueller et. al. [4] |
| cce\_3127 | cce\_4412 | 0.39391 | Mueller et. al. [4] |
| cce\_3127 | cce\_3517 | 0.35169 | Mueller et. al. [4] |
| cce\_3127 | cce\_2967 | 0.32985 | Mueller et. al. [4] |
| cce\_3127 | ho2 | 0.31086 | Mueller et. al. [4] |
| cce\_3127 | ndhF1 | -0.27927 | Mueller et. al. [4] |
| cce\_3127 | metS | -0.30625 | Mueller et. al. [4] |
| cce\_3127 | cce\_1482 | -0.34607 | Mueller et. al. [4] |
| cce\_3127 | isiB2 | -0.26273 | Mueller et. al. [4] |
| cce\_3127 | cce\_1155 | -0.29681 | Mueller et. al. [4] |
| cce\_3127 | sigE | -0.28941 | Mueller et. al. [4] |
| sigB | ntcA | 1 | McDermott et. al. [3] |
| sigD | pheA | -0.32967 | Mueller et. al. [4] |
| sigD | cce\_4412 | 0.59383 | Mueller et. al. [4] |
| sigD | chlH1 | 0.56801 | Mueller et. al. [4] |
| sigD | psaL1 | -0.44247 | Mueller et. al. [4] |
| sigD | cce\_3941 | -0.11331 | Mueller et. al. [4] |
| sigD | cce\_3870 | 0.05658 | Mueller et. al. [4] |
| sigD | psb27 | 0.24853 | Mueller et. al. [4] |
| sigD | ndhD2 | 0.33191 | Mueller et. al. [4] |
| sigD | cce\_3518 | -0.32317 | Mueller et. al. [4] |
| sigD | cce\_3107 | -0.6131 | Mueller et. al. [4] |
| sigD | glgP1 | 0.39017 | Mueller et. al. [4] |
| sigD | rub1 | -0.21166 | Mueller et. al. [4] |
| sigD | cce\_0571 | -0.31863 | Mueller et. al. [4] |
| patB | rpaA | -1 | McDermott et. al. [3] |
| patB | ntcA | -1 | McDermott et. al. [3] |
| patB | nifH | 1 | McDermott et. al. [3] |
| patB | nifD | 1 | McDermott et. al. [3] |
| patB | nifK | 1 | McDermott et. al. [3] |

The algorithm used to infer an edge between a transcription factor and target gene along with the literature source reference is given below:

|  |  |  |
| --- | --- | --- |
| Literature Source Number | Literature Source | Algorithm Used |
| 1 | Vinh et. al. [2] | GlobalMIT+ |
| 2 | McDermott et. al. [3] | CLR and Inferelator |
| 3 | Mueller et. al. [4] | An Optimization Framework |

The mathematical details of the algorithms used to predict the network is presented below:

### Global MIT+ (Algorithm used by Vinh et. al. [2] to infer the edges in the above network)

The focus of this algorithm is on the problem of learning the globally optimal structure for the first-order Markov stationary Dynamic Bayesian Network (DBN). This form of learning is referred to as the search + score approach, in which a scoring function is specified to assess the goodness-of-fit of a DBN given the data, and a search procedure to find the optimal network based on this scoring metric.

The scoring function used in this algorithm is called the MIT score, originally used for learning Bayesian Networks, which can then be adapted to the DBN case. Briefly speaking, under MIT the goodness-of-fit of a network is measured by the total mutual information shared between each node and its parents, penalized by a term which quantifies the degree of statistical significance of this shared information.

#### Mutual Information between discrete variables

Mutual Information gives the measurement of the reduction in uncertainty of parts of the outcome of a system given measurement of the other parts. Entropy given by the formula stated below is an ideal measure of uncertainty in the system.

Here, H(X) is the entropy of the random variable X, P(X) is the probability of its occurrence. For two variables X and Y, the joint entropy is given by the following formula:

Here, H (X, Y) is the joint entropy of the random variables X and Y, P (X, Y) is the joint probability of their occurrence. Mutual Information can be written in terms of the entropies and the joint entropy of 2 random variables as

or

or

#### Learning Bayesian Network using MI scoring function

The method for learning Bayesian networks based on the score + search paradigm can be expressed as follows:

Given a variable with corresponding discrete states and a complete training dataset of instances, find a Directed Acyclic Graph such that:

Given a DAG G, the mutual information between a variable and its parent can be given as . Let be the number of parent variables of

The conditional mutual information property says that:

MI (X, Y, W|Z) = MI (X, Y|Z) + MI (X, W|Z, Y)

We can iteratively apply the above property in our case to get

The elements in this decomposition can be interpreted as starting with an empty set of parent of, we have first included the arc and the degree of dependence between the variables is. We then insert the arc and as is already a parent of, the degree of dependence is. We continue to do so until we insert the last arc. Next, we want to determine how the addition of an arc can represent a statistically significant increase in mutual information. approximates to a distribution with the appropriate degrees of freedom. If we fix the confidence level to and determine the value of such that this represents a statistical test of conditional independence. Thus, we can express the scoring functions as

#### MI scoring function for Dynamic Bayesian Networks (DBN)

For Dynamic Bayesian Networks, the mutual information is calculated between a parent set and it's child which is shifted by 1-unit in time as per the first order Markov assumption denoted by. The number of effective observations for separate time series is equal to. The MIT score becomes:

### Context Likelihood of Relatedness (Algorithm used by McDermott et. al. [3] to infer the edges in the above network)

#### CLR Introduction

The Context Likelihood of Relatedness (CLR) algorithm is an extension of the relevance network approach for identifying transcriptional regulatory interactions. The original relevance network method used mutual information for scoring the similarity between the expression levels of two genes in a set of microarrays. A gene and a transcription factor are predicted to interact if the mutual information between the expression levels of the gene and its potential regulator is above some set threshold.

#### Mutual Information between continuous variables

The above formulas to calculate MI are for discrete random variables. For continuous random variables there are 2 approaches.

We can take the easy way out i.e. binning the data into M discrete intervals . For experimental data consisting of N measurements of a variable an indicator function counts the number of data points within each bin. The probabilities are then estimated based on the relative frequency of occurrence

with

For two variables, the joint probabilities are calculated from a multivariate histogram. It has been suggested to adaptively choose the sizes of the bins so that each bin has a uniform distribution of points. However, in this approach, each data point is assigned to one and only one bin. For datapoints near the boundary of the bin, small fluctuations may shift these to neighboring bins. The positions of the borders can therefore strongly affect the resulting mutual information. To overcome the drawbacks, datasets are assigned to multiple bins simultaneously. Thus, the indicator function is extended to be a set of polynomial B-spline functions.

#### B-Spline Functions

The first step towards replacing the indicator function with the B-spline function is the calculation of the knot vectors If the number of bins that each variable is assigned to is M and the degree of the polynomial of the spline is k, then the knot vectors will range from 0 to M-1+k. The knot vector values are as follows:

Next it is required to define z which is the normalized form of the variable values and will be used to calculate the B function.

Then the weights of the values for every bin is calculated using the B-spline function which is given by the following formula,

and the following recursive formula

Once the weighting coefficients for each x value are calculated, we determine the probability for each bin from this formula:

The joint probability function is given below:

With the probability and joint probability of the variables defined, we can determine the entropy of each variable, joint entropy of two variables and finally the mutual information between them.

#### CLR filtering step

After computing the mutual information between regulators and their potential target genes, CLR calculates the statistical likelihood of each mutual information value within its network context. The algorithm compares the mutual information between a transcription factor/gene pair to the ‘‘background’’ distribution of mutual information scores for all possible transcription factor/gene pairs that include either the transcription factor or its target. The most probable interactions are those whose mutual information scores stand significantly above the background distribution of mutual information scores.

The background distribution is constructed from two sets of MI values: , the set of all the mutual information values for gene i (in row or column i ), and the set of all the mutual information values for gene j (in row or column j). Because of the sparsity of biological regulatory networks, most MI scores in each row of the mutual matrix represent random background MI (e.g., due to indirect network relationships). We approximate this background MI as a joint normal distribution with and as independent variables, which provides a reasonable approximation to the empirical distribution of mutual information. Thus, the final form of our likelihood estimate becomes

where and are the z-scores of from the marginal distributions, and is the joint likelihood measure.

### Inferelator (Algorithm used by McDermott et. al. [3] to infer the edges in the above network)

Here, we assume that the expression level of a gene, or the mean expression level of a group of co-regulated genes y, is influenced by the level of N other factors in the system: X =. Methods for selecting which of these factors are the most likely regulators, among all possible regulatory influence factors are described below.

Here, is a set of functions of the regulatory factors X. The coefficients describe the influence of each element of Z, with positive coefficients corresponding to inducers of transcription and negative coefficients to transcriptional repressors. The choice amounts to the simple weighted linear combination of influencing factors . Various functional forms can be adopted for the function g, called the 'nonlinearity' or 'activation' function for artificial neural networks, and the 'link' function in statistical modeling. The function g often takes the form of a sigmoidal, or logistic activation function

Given our model formulation, for selecting subsets of predictors we adopt the L1 shrinkage or the LASSO:

subject to:

where is the ordinary least square estimate of .The shrinkage parameter t can range from 0 to 1.

### An Optimization Framework (Algorithm used by Mueller et. al. [4] to infer the edges in the above network)

An optimization algorithm was used to construct the network. The network of regulatory influences for each organism was generated by iteratively determining for each gene cluster the maximum amount of experimental expression change that can be accounted for a given number of regulatory influences. J, K and T denote the total number of gene clusters, TF clusters, and time points, respectively. The experimental expression level of gene cluster j at time point t is represented as and the experimental expression level of TF cluster k at time point t is denoted as . is the interaction coefficient that describes the influence of TF cluster k on gene cluster j. Gene expression was assumed to be a linear additive contribution of the set of influencing TFs. This can be postulated for each gene cluster j and time point t as shown in:

Backward finite difference was used to approximate the rate of expression term (i.e., left-hand side) in the mixed integer linear program formulation shown above. Solving this formulation identifies the interaction coefficients that minimize the amount of change in gene cluster expression not explained by the regulatory influences. The formulation was solved iteratively for every gene cluster and a range of maximum regulatory influences.

s.t.

The slack variables and represent the positive and negative deviations from experimental measurements. The binary variables, , are used to restrict the number of regulatory influences per gene cluster. If a transcription factor in TF cluster k is in gene cluster j then a regulatory interaction will be imposed for that TF cluster gene cluster pair by fixing to one for that pair. In order to determine the number of regulatory influences per cluster, the percentage of experimental expression change accounted for by the network was calculated for the range from one to 11 regulatory influences. For a given number of regulatory influences this percentage of explained expression variance is defined as the difference between the error (i.e., the sum of the slack variables) with zero regulatory influences and the given number of regulatory influences, divided by the error with zero influences. The number of regulatory influences was chosen for each cluster as the minimum number of influences that brought the percentage of explained expression change above 50%.

## Functional evidence of some of the transcription factors found in other organisms

Among the 17 Transcription Factors listed above, some of them have very high sequence similarity with genes in other organisms and they may be homologs while others are uncharacterized or have unknown function. A detailed description of the transcription factors is given below:

1. cce\_0678: According to [3], this gene is described as an “uncharacterized two-component regulator(cce\_0678) that bears a strong resemblance to cce\_0298 (rpaA). The exact function of these proteins is not known, but the parallelism of cce\_0678 for photosynthesis and cce\_0298 (rpaA) for nitrogen metabolism is striking. These two regulators may be positively or negatively regulated by similar input signals and work in parallel to favor either photosynthesis or nitrogen fixation.” As evident in the above network, this gene activates the RuBisCO complex which plays a central role in Carbon Fixation during photosynthesis.
2. sigE: According to [5], “In Synechocystis sp. PCC 6803, SigE specifically contributes to the expression of genes related to sugar catabolism and photosynthesis, and is regulated by the circadian system at the transcript and protein levels. Therefore, SigE may be a σ factor controlling the balance of carbon and nitrogen metabolism with a rhythmic expression that peaks at 24-hour intervals according to the upcoming night”. As seen in the network, SigE interacts with various genes involved in nitrogen fixation like nifZ or glycolysis like cce\_3941 which according to it’s functional annotation given in [1] is an alpha amylase that hydrolyses alpha bonds of large, alpha-linked polysaccharides, such as starch and glycogen, yielding glucose and maltose.
3. cce\_4141: This is a transcription factor with an unknown function. However, it interacts with another transcription factor, cce\_3298 which in turn interacts with both sigB and ntcA that are known to play a role in the expression of nitrogen metabolism related genes as described below.
4. kaiA: kaiA is part of the circadian clock protein complex. According to the network, it interacts with cce\_4751 which has been annotated as a Two-component hybrid sensor and regulator.
5. sasA: This is a kaiC-interacting sensory histidine kinase that is necessary to sustain robust oscillation in Synechococcus sp. 7942 and is a part of a two-component regulatory system according to [6]. It is seen in the network that sasA interacts with kaiC2.
6. cce\_3127: This is a transcriptional regulator from the Fur family according to its functional annotation. Fur-type transcriptional regulators are involved in regulating the uptake and homeostasis of several metal ions and related functions in Synechococcus sp. PCC 7002 according to [7].
7. cce\_1520: There is not enough information to say anything about this transcription factor.
8. kaiB1: Part of the circadian clock protein complex kaiABC.
9. patB: According to [3], “The regulator ntcA is thought to play a central role in regulation of heterocyst formation in Anabaena in response to nitrogen starvation. Additionally, patB is known to be specifically upregulated late in heterocyst formation in response to nitrogen starvation, is a member of a conserved core set of genes along with nitrogenase and is thought to be sensitive to redox state. Finally, rpaA is a member of a two-component system involving the sasA gene product that is closely coupled to the KaiABC circadian oscillatory system and regulates functions involved in energy transfer from photosystem to the phycobilisome. Therefore, it appears that there may be a feedback loop between patB and ntcA, which seem to play opposing roles in Cyanothece51142”.
10. sigD: According to [5], “In 2003, it was reported that the light-induced σ factor PCC 6803 SigD specifically recognized the promoters of photosynthetic genes, psbA2and psbA3, and contributed to their light-induced transcription. Since then, it has been shown that SigD also contributes to the transcription of the light-induced photosynthetic genes cpcBADC, petBD, and psaAB. This demonstrates a universal function of SigD for light-induced transcription in cyanobacteria”. In the above network. sigD interacts with psaL1, a Photosystem I reaction center and psb27, a photosystem II lipoprotein which are both involved in photosynthesis.
11. cce\_3378: This transcription factor has been annotated as a part of a two-component regulator and is expressed during the light to dark transition phase as obtained from the supplementary material in [3]. It was also classified as one of the top 25 topological bottlenecks in the network developed by [3]. From the network it is seen to affect multiple target genes including the kaiB3 circadian clock protein.
12. sigB: According to [5], "In *Synechocystis* sp 6803, SigB contributes to the nitrogen responsive transcription of glnB and other ntcA dependent genes". The above network predicts that sigB interacts with ntcA which is known to be a global nitrogen regulator and is in accordance with the above statement.
13. rpaA: This Transcription Factor has been established to be a Master Regulator in Synechococcus elongatus PCC 7942 by [8]. Through chromatin immunoprecipitation with high-throughput sequencing (ChIP-seq) and in vitro DNase I footprinting, they have shown that phosphorylated RpaA binds directly to >100 locations in the genome, including the promoter of kaiBC. The above network also predicts rpaA to interact with multiple transcription factors and target genes that are involved in photosynthesis or nitrogen fixation pathways.
14. kaiC2: kaiC2 is predicted to be the primary clock protein as per [2]. It has been shown in Synechococcus elongatus PCC 7942 that this is one of the core clock proteins along with kaiA and kaiB. Cyanothece has 2 copies of kaiC.
15. sigJ: According to [5], in Anabaena sp. PCC 7120, SigJ is a key regulator of desiccation tolerance and regulates the synthesis of extracellular polysaccharide. Accommodating dehydration is important for cell survival, since drying causes cell lysis and damage to nucleic acids, proteins, and membranes.
16. cce\_3298: This is an unknown protein. However, it interacts with both sigB and ntcA which are known to are known to play a role in the expression of nitrogen metabolism related genes as given above.
17. kaiB3: This is part of the circadian clock protein complex kaiABC.

## Conclusion

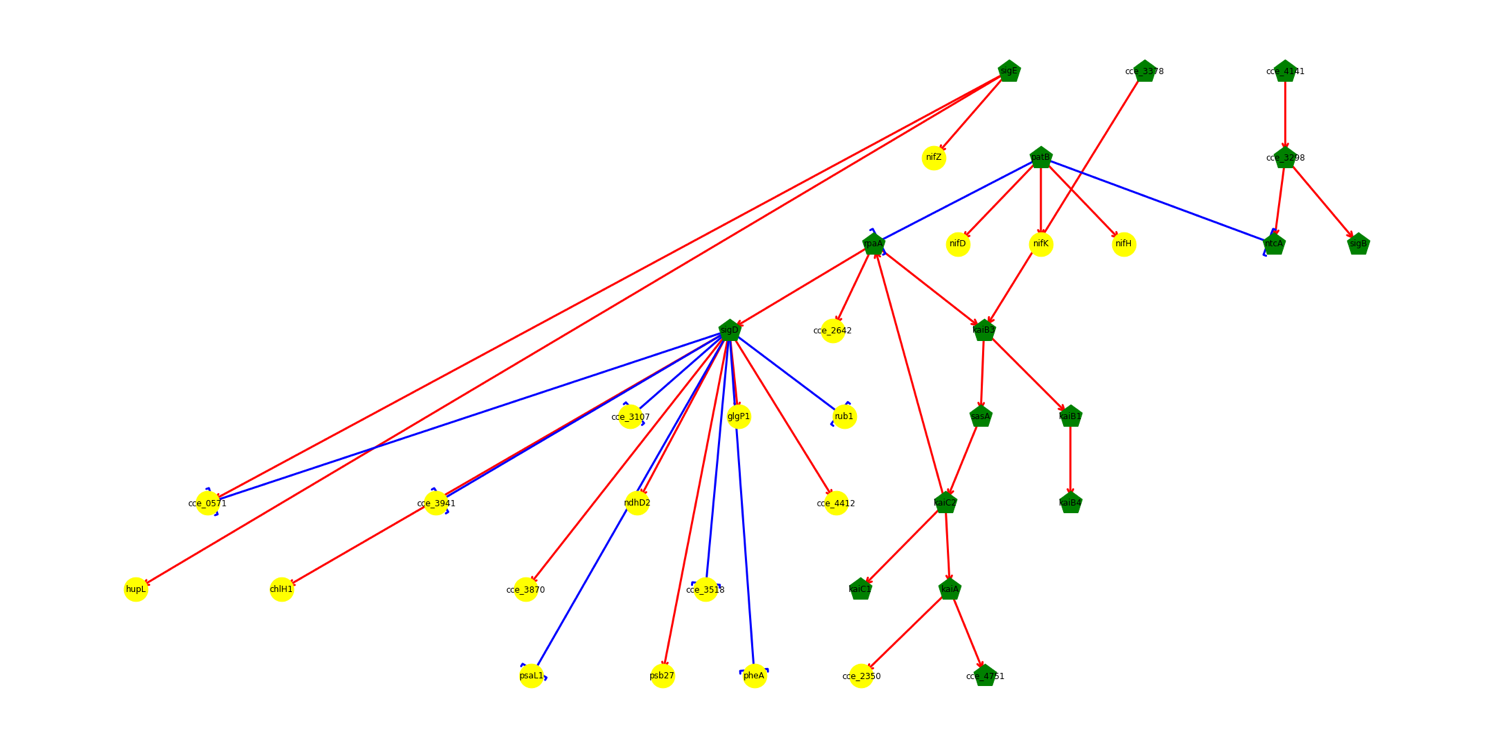
The gene regulatory network shown in Figure 1 above has highlighted several potential key regulators that may take part in the downregulation of photosynthesis and carbon fixation and the upregulation of nitrogen fixation and respiration during the light/dark transition. While some of them like rpaA has been established to play the role of a master regulator in other cyanobacteria like Synechococcus elongatus PCC 7942 by [8] others like cce\_3378 or cce\_0678 has unknown functions and they have not been studied well. Based on the network and the above functional evidence the following transcription factors that have the potential to play key role in the transition phase are:

1. rpaA: Markson et. al. [8] has shown that RpaA plays the role of a master regulator in *Synechococcus elongatus* PCC 7942. Deletion of rpaA arrests the cells in a subjective dawn-like transcriptional state and circadian gene expression oscillations is absent genome wide in a rpaA mutant. Moreover, through chromatin immunoprecipitation with high-throughput sequencing (ChIP-seq) they have shown that phosphorylated RpaA binds directly to >100 locations in the genome including the promoter of KaiBC. Additionally, they have also shown that overexpression of a phosphomimetic mutant of RpaA is sufficient to drive cells from the subjective dawn to the subjective dusk gene expression state demonstrating that RpaA is a global regulator of the circadian output in *Synechococcus elongatus* PCC 7942. In the network given above, it has been predicted that rpaA interacts with KaiB3 and KaiC2 as well as other important transcription factors like sigD and ntcA.
2. patB: PatB has been identified previously as a promising candidate to control the nif cluster expression [4] and from the above network, it has been predicted that patB interacts with nifHDK operon which takes part in nitrogen fixation. It also inhibits rpaA in the given network which has been shown to be a master regulator in *Synechococcus elongatus* PCC 7942 as stated above.
3. cce\_4141 and cce\_3298: Although not much is known about these 2 genes, the gene regulatory network predicts that they interact in a very interesting manner. Firstly, cce\_3298 has been predicted to interact with both sigB and ntcA which are known to take part in nitrogen regulation as given above. Secondly, cce\_4141 has been shown to affect cce\_3298 and other target genes which are involved in either the glycogen breakdown pathway like glgP1 or in the amino acid biosynthesis pathway like gdhA.
4. cce\_3378: This is another unknown transcriptional factor that is expressed during the light to dark transition phase, was classified as one of the top 25 network topological bottleneck by [3] on the basis of betweenness centrality and node centrality calculated from the network topology and can be seen to affect multiple target genes along with the circadian clock protein kaiB3 in the network given above.
5. cce\_0678: cce\_0678 has been shown to interact with the RuBisCO complex formed by rbcS and rbcL genes, that plays a central role in carbon fixation which is closely linked with photosynthesis, by McDermott et. al. [3]. They have drawn comparison between cce\_0678 and rpaA and stated that there is a strong resemblance between the way these genes function based on their inferred network. Although not much is known about this gene, if it is indeed functionally parallel to rpaA and is regulated by the same input signals as they have claimed on the basis of their inferred network, then this is something worth looking into.
6. sigB, sigD and sigE: These sigma factors are known to play important roles in other cyanobacteria as stated in the above section. SigB for example is nitrogen starvation responsive and is known to be specifically expressed in response to nitrogen starvation in *Synechococcus* sp. PCC 7002 and in *Synechocystis* sp. 6803 according to Imamura et. al. [5]. SigD on the other hand is said to contribute to the light induced transcription of photosynthetic genes in *Synechocystis* sp. 6803. SigE has been predicted to be the sigma factor that control sugar catabolism and is regulated by the circadian system. These types of interactions are seen in the network given above as mentioned in the previous section. Therefore, these sigma factors may play an important role in the transition phase in *Cyanothece* as well.

## Additional Subnetworks

### Circadian

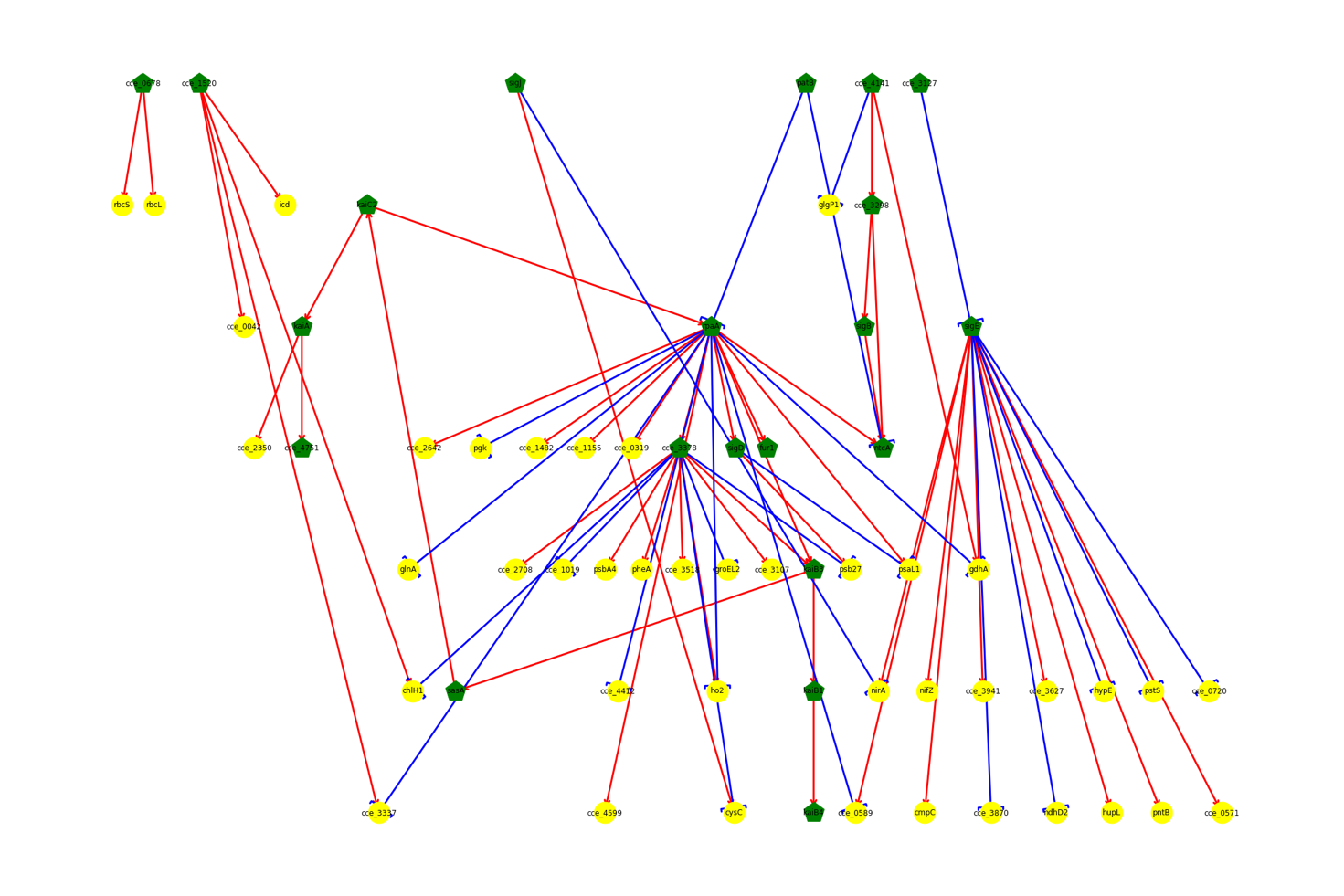
Only those transcription factor and target gene that are predicted to be under circadian rhythm as classified by [3] and their interactions with other transcription factors and target genes are in this network



*To better visualize the network, an image file named cGRN.png has also been provided along with this report.*

### Diurnal

Only those transcription factor and target gene that are predicted to be under diurnal rhythm as classified by [3] and their interactions with other transcription factors and target genes are in this network



*To better visualize the network, an image file named dGRN.png has also been provided along with this report.*

# References

|  |  |
| --- | --- |
| [1] | E. A. Welsh, M. Liberton, J. Stöckel, T. Loh, T. R. Elvitigala, C. Wang, A. Wollam, R. S. Fulton, S. W. Clifton, J. M. Jacobs, R. Aurora, B. K. Ghosh, L. A. Sherman, R. D. Smith, R. K. Wilson and H. B. Pakrasi, "The genome of Cyanothece 51142, a unicellular diazotrophic cyanobacterium important in the marine nitrogen cycle," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 105, no. 39, pp. 15094-15099, 2008. |
| [2] | N. X. Vinh, M. Chetty, R. L. Coppel, S. B. Gaudana and P. P. Wangikar, "A model of the circadian clock in the cyanobacterium Cyanothece sp. ATCC 51142," *BMC Bioinformatics,* vol. 14, no. 2, pp. 1-9, 2013. |
| [3] | J. E. McDermott, C. S. Oehmen, L. A. McCue, E. A. Hill, D. M. Choi, J. Stöckel, M. Liberton, H. B. Pakrasi and L. A. Sherman, "A model of cyclic transcriptomic behavior in the cyanobacterium Cyanothece sp. ATCC 51142," *Molecular BioSystems,* vol. 7, no. 8, pp. 2407-2418, 2011. |
| [4] | T. J. Mueller, E. A. Welsh, H. B. Pakrasi and C. D. Maranas, "Identifying Regulatory Changes to Facilitate Nitrogen Fixation in the Nondiazotroph Synechocystis sp. PCC 6803," *ACS Synthetic Biology,* vol. 5, no. 3, pp. 250-258, 2016. |
| [5] | S. . Imamura and M. . Asayama, "Sigma Factors for Cyanobacterial Transcription," *Gene regulation and systems biology,* vol. 3, no. 3, pp. 65-87, 2009. |
| [6] | N. M. O. T. K. R. S. C. S. M. K. T. I. H. Takai N, "A KaiC-associating SasA-RpaA two-component regulatory system as a major circadian timing mediator in cyanobacteria," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 103, no. 32, p. 12109–14, . |
| [7] | M. Ludwig, T. C. Tiing, Y. C. Chyue and D. A. Bryant, "Fur-type transcriptional repressors and metal homeostasis in the cyanobacterium Synechococcus sp. PCC 7002". |
| [8] | J. S. Markson, J. S. Markson, J. R. Piechura, J. R. Piechura, A. M. Puszynska, A. M. Puszynska, E. K. O'Shea and E. K. O'Shea, "Circadian Control of Global Gene Expression by the Cyanobacterial Master Regulator RpaA," *Cell,* vol. 155, no. 6, pp. 1396-1408, 2013. |