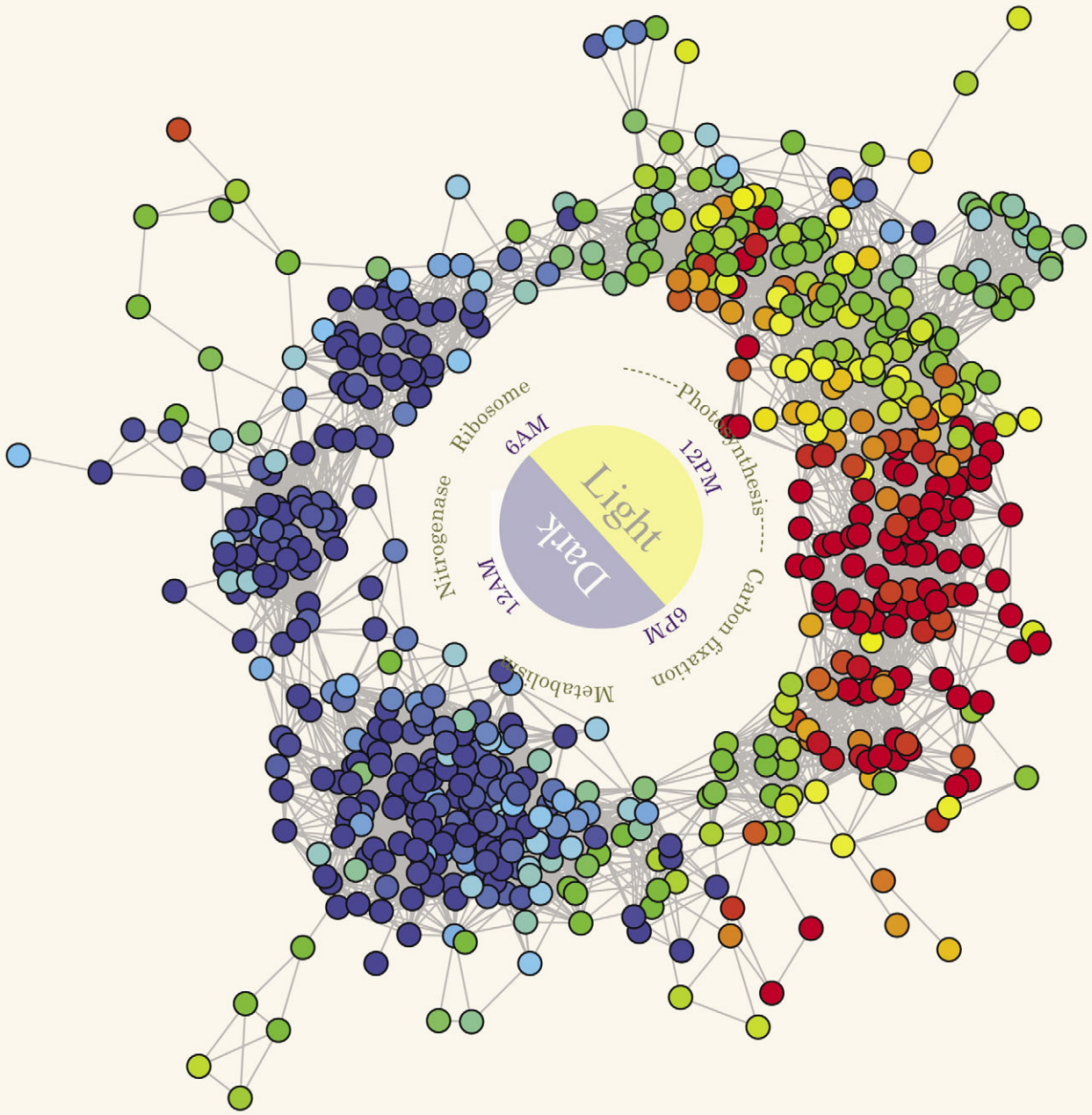
Molecular BioSystems

www.molecularbiosystems.org

Volume 7 | Number 8 | 1 August 2011 | Pages 2333–2526



ISSN 1742-206X

RSC Publishing

PAPER

Jason E. McDermott *et al.*

A model of cyclic transcriptomic behavior in the cyanobacterium *Cyanothece* sp. ATCC 51142

Molecular BioSystems

Cite this: *Mol. BioSyst.*, 2011, **7**, 2407–2418

www.rsc.org/molecularbiosystems

PAPER

A model of cyclic transcriptomic behavior in the cyanobacterium *Cyanothece* sp. ATCC 51142†

Jason E. McDermott,*[a] Christopher S. Oehmen,[a] Lee Ann McCue,[a] Eric Hill,[b] Daniel M. Choi,[a] Jana Stöckel,[c] Michelle Liberton,[c] Himadri B. Pakrasi[c] and Louis A. Sherman[d]

Received 6th January 2011, Accepted 28th May 2011
DOI: 10.1039/c1mb05006k

Systems biology attempts to reconcile large amounts of disparate data with existing knowledge to provide models of functioning biological systems. The cyanobacterium *Cyanothece* sp. ATCC 51142 is an excellent candidate for such systems biology studies because: (i) it displays tight functional regulation between photosynthesis and nitrogen fixation; (ii) it has robust cyclic patterns at the genetic, protein and metabolomic levels; and (iii) it has potential applications for bioenergy production and carbon sequestration. We have represented the transcriptomic data from *Cyanothece* 51142 under diurnal light/dark cycles as a high-level functional abstraction and describe development of a predictive *in silico* model of diurnal and circadian behavior in terms of regulatory and metabolic processes in this organism. We show that incorporating network topology into the model improves performance in terms of our ability to explain the behavior of the system under new conditions. The model presented robustly describes transcriptomic behavior of *Cyanothece* 51142 under different cyclic and non-cyclic growth conditions, and represents a significant advance in the understanding of gene regulation in this important organism.

Introduction

Organisms from cyanobacteria to humans display rhythmic behavior closely linked to circadian and diurnal cycles. Many systems utilize a complex interplay between circadian rhythms that provide internal temporal cues, and the diurnal cycle, which often serves to entrain the circadian machinery using external inputs.[1] Organisms that rely on photosynthesis have evolved complicated systems for regulation of diurnal rhythms to deploy photosynthetic machinery in response to light and circadian mechanisms to ensure that the organism is ready for the light period.[2–5] The interaction of environmental cues, such as light and temperature, and the activity of the circadian clock is an area of intense study.[2–4,6–11]

Cyanothece sp. ATCC 51142 (here, *Cyanothece* 51142) is an important bacterium in benthic environments, both as a fixer of atmospheric nitrogen and as a photosynthetic primary producer that evolves O_2.[12] The two processes, however, are incompatible, as the nitrogenase enzyme is extremely sensitive to oxygen. This challenge in diazotrophic cyanobacteria is usually met with the formation of specialized heterocysts in filamentous cyanobacteria such as *Anabaena* and *Nostoc*,[13] which provide a spatial separation of the two pathways. *Cyanothece* 51142, on the other hand, separates these pathways temporally.[14] It does so using a robust "clocking" mechanism that divides central metabolic processes diurnally, undergoing photosynthesis during light cycles and nitrogen fixation during the dark.[5] The tight organization of metabolic processes in *Cyanothece* 51142 were also found to be clocked through a circadian rhythm and a transcriptional analysis over light/dark (LD) cycles showed tight clustering of photosynthesis-related, nitrogenase-related, and respiration-related transcripts; the inferred network based only on statistical relatedness resulted in a functional "clock" of activity.[5,15]

We and others have undertaken global transcriptomic studies of cyanobacteria[3,5,16] during light-dark cycles to discover genes that cycle in response to circadian, diurnal, or other cues. These studies have identified regulatory interactions and defined functional modules which are temporally distinct by identifying individual genes that cycle and finding correlations between genes with similar functions. We have used networks of inferred regulatory relationships to visualize and analyze complicated transcriptomic data.[5,15] Topological analysis of

[a] *Computational Biology and Bioinformatics Group, Pacific Northwest National Laboratory, MSIN: J4-33, 902 Battelle Boulevard, PO Box 999, Richland, WA 99352, USA. E-mail: Jason.McDermott@pnl.gov; Fax: 509-372-4720; Tel: 509-372-4360*
[b] *Microbiology, Pacific Northwest National Laboratory, Richland, WA 99352, USA*
[c] *Department of Biology, Washington University, St. Louis, MO 63130, USA*
[d] *Department of Biological Sciences, Purdue University, West Lafayette, IN 47907, USA*
† Electronic supplementary information (ESI) available. See DOI: 10.1039/c1mb05006k

these networks can be used to provide biological insight into important genes in the system. Proteins that are highly central in protein-protein and regulatory interaction networks, so-called bottlenecks and hubs, were shown more likely to be important to the system in several studies.[17–19] We have used a similar approach to analyze networks from inferred transcriptomics[20,21] or global proteomics measurements,[22] and to identify putative mediators of transitions between system states.

A number of mathematical models of cyanobacterial behavior have been developed and largely focused on aspects of the circadian clock machinery.[23,24] A recently described model expands on this approach to describe multiple aspects of the system at both abstract and molecular levels, using an agent-based modeling approach.[25] Metabolic models based on existing knowledge about enzymatic reactions and a set of simplifying assumptions have also been developed for cyanobacterial species.[26] These kinds of bottom-up models can be limited in their ability to make predictions about metabolic functions not included in the construction of the model or about global patterns of transcription. In contrast top-down approaches strive to generate useful models directly from high-throughput data generated for the system, with little reliance on existing knowledge. We[2,5,15,27] and others[28,29] have used various methods to infer networks of regulatory associations from high-throughput data, shedding light on the overall organization of the transcriptional programs of cyanobacteria at a high abstraction level.

Recent developments in this area have produced computational models from high-throughput data that are predictive of the global transcriptional regulatory program of the organism. Bonneau, *et al.*, developed a method to infer a parsimonious set of regulatory influences that accurately describe the transcriptional behavior of a set of targets, co-regulated sets of genes in *Halobacterium*.[30] The method uses gene expression profiles from both equilibrium and time course experiments to fit ordinary differential equations (ODEs) and selects the minimal set of most informative regulatory influences for each target cluster. Models such as these rely on a simplified version of the system based on clustering of genes with similar behavior into functional modules to identify and parameterize regulatory influences. These kinds of models can be used to predict the global behavior of a system using measurements from a small number of regulators or other input parameters, to predict the behavior of the system at a future time point, and to formulate predictions based on *in silico* manipulation of the model, for example regulator knock-downs or variance of environmental conditions.

In the current study we report the development of a predictive model of cyclic behavior in *Cyanothece* 51142 using a previously published method, the Inferelator.[30] The model is based on a set of transcriptional experiments that are focused on investigating diurnal and circadian processes in this organism. We report that the model can accurately predict the behavior of the system when validated on independent data. We found that topology derived from co-expression networks was correlated with gene conservation and that including topological bottlenecks as potential regulators improves the performance of the predictive model. Functional modules, *i.e.* targets of the inference process, were defined using an iterative process of modeling. We found that the behavior of portions of the metabolic network representing important metabolic processes, *e.g.* nitrogenase and ribulose-1,5-bisphosphate carboxylase oxygenase (RuBisCO), could be accurately predicted using our models. Finally, we show that the model trained on cyclic time course data is capable of predicting expression dynamics in an acyclic validation time course experiment following *Cyanothece* 51142 in low oxygen conditions. The models we describe represent an important step forward in the systems biology of photosynthetic cyanobacteria and provide a large number of insights into important biological processes under cyclic regulation.

Results and discussion

Network analysis of *Cyanothece* 51142 dynamics

Previously we have used co-expression networks to determine functional modules and represent dynamic processes of *Cyanothece* 51142 at a transcriptional level.[5,15] A powerful tool to assess the importance of individual genes in regulation of the system is network topology.[20,22,31] Since there is little known about the regulatory structure of *Cyanothece* 51142, we wanted to ascertain if approaches based on network topology could identify important regulators of system dynamics.

We first inferred networks between genes using Pearson correlation or the context likelihood of relatedness (CLR) method.[32] Each method uses the similarities between expression profiles of genes to determine relationships that can represent regulation (gene A regulates gene B), co-regulation (genes A and B are both regulated by gene C) or co-expression (genes A and B expressed at the same time). The Pearson correlation network forms a 'wreath' structure that is temporally ordered (see the ESI†) and is shown in Fig. 1. The colors in Fig. 1 indicate membership of clusters derived from the full ensemble of time course data. Therefore some of the clusters in this network, which was based only on the 12 h LD data, appear discontinuous. We then identified topological bottlenecks by calculating the betweenness centrality of all the genes in networks. This measure is based on the number of times that a gene is used by the shortest paths between all other pairs of genes in the network. To assess the importance of the genes in the network we used the evolutionary conservation of the gene (see Methods). In general, conserved genes may be more important to a system, because they represent functions that have been selected over evolutionary time. Based on previous studies,[17,18,20] we considered the 20% of the genes in the network with the highest betweenness values as bottlenecks. We found that bottlenecks were significantly more likely to be present in the closely related cyanobacterium *Synechocystis* sp PCC 6803, in *Escherichia coli*, or in the plant *Arabidopsis thaliana*, and shared in all photosynthetic organisms in general (*Synechocystis*, *Anabaena*, and *A. thaliana*) than both the average of other genes in *Cyanothece* 51142 and other cyclic genes in the network (Fig. S1, ESI†) (p value < 0.01 by Chi-square test). The top topological bottlenecks from this network are listed in Table 1, which also shows their

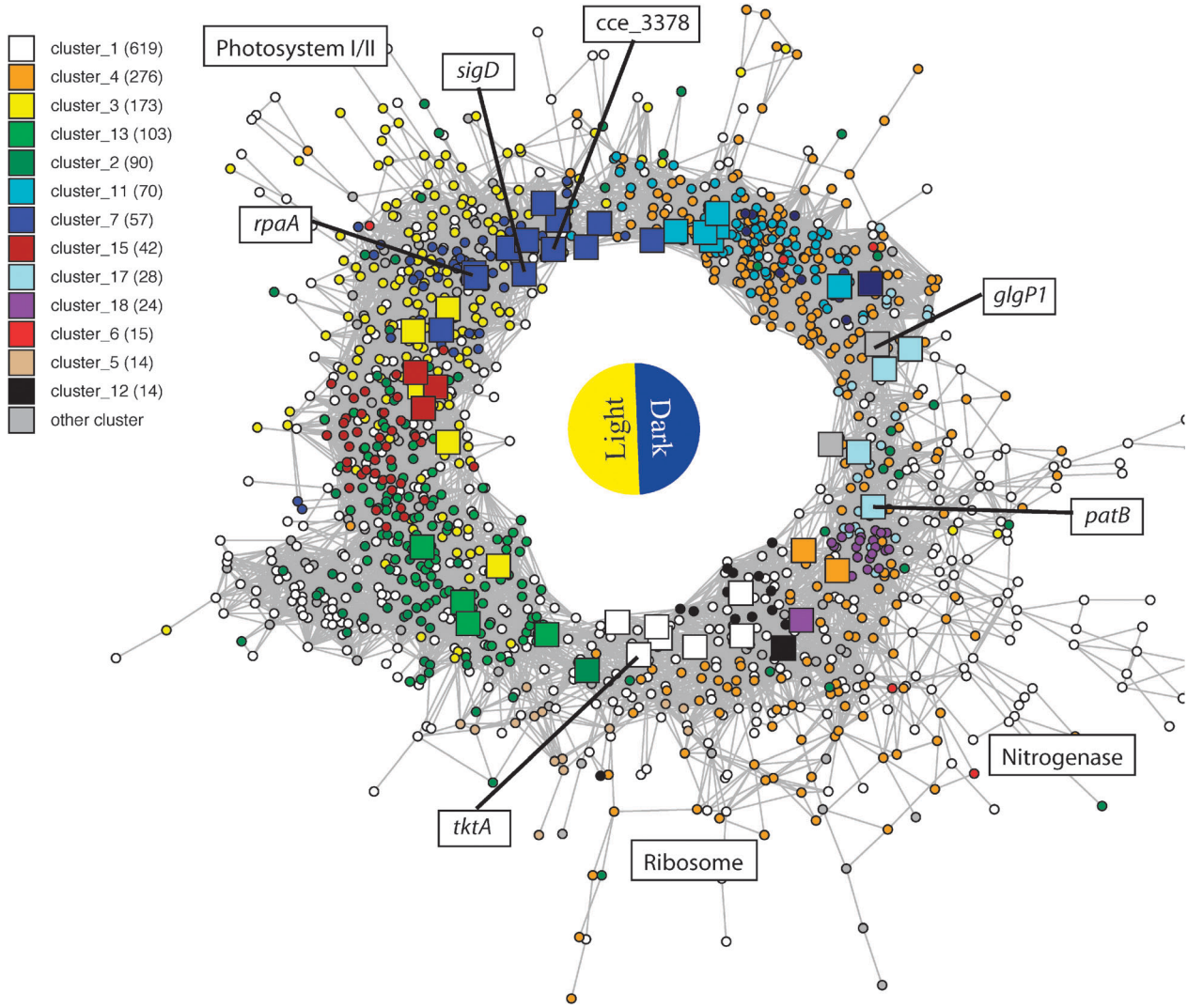


Fig. 1 Topology of the cyclic wreath network for *Cyanothece* 51142 transcription. A co-expression network of the transcriptomic profiles from *Cyanothece* 51142 genes under 12 h LD cycles was constructed. The temporal ordering of gene expression in *Cyanothece* 51142 is represented in the network by the location of a gene (node) at its peak expression. Colors represent clusters identified by hierarchical clustering of the complete dataset that includes three other time course experiments (see legend and Table 2). Topological bottlenecks were identified from the network, and the top 5% shown as squares in the network. Additionally, the temporal location of functional groups and several known regulators of systems transitions are labeled.

classification as diurnal or circadian using conservative criteria, the cluster they belong to (see Table 2) and their peak expression time (transition). We noted the presence of several key regulators known to be mediators of transitions between system states in closely related systems (shown in Fig. 1) and these are discussed below. Our hypothesis, based on our results[20,22] and those of others,[33] is that these bottlenecks represent mediators of transitions between different biological states in the system. That is, they participate in the function of two (or more) functional modules, and may represent points of control when the system moves from one state to another.

Roles of topological bottlenecks in systems transitions

Functional enrichment of bottlenecks shows that they are enriched in a number of processes previously identified as important for *Cyanothece* 51142 (Table S1, ESI†), although this enrichment was close to the significance threshold due to the small number of bottlenecks examined. The top bottlenecks include several genes that are known to play a role in transitions between system states. A *patB* homolog directly precedes the nitrogenase cluster in our network and is known to be involved in the induction of nitrogenase activity in *Anabaena*.[34] We have previously shown that *sigD*, a RNA polymerase sigma factor, which has a gene expression peak at the end of the light period spanning into the early dark, plays a critical role in this transition.[35] The *rpaA* gene, a transcriptional regulator, peaks late in the day, concurrent with *sigD*. It is known from mutational studies in *Synechocystis* that the *rpaA*/*sasA* two component regulatory system is a major component of the circadian timing system and associates with the phosphorylated form of the KaiC protein.[36] Two members of the OPP pathway are found in the list of top bottlenecks. The transketolase A (*tktA*) gene was shown to be upregulated

Table 1 Top 25 topological bottlenecks

ID	Rank	Name	Cyclicity	Description	Cluster	Transition
cce_4095	1		circadian	unknown; contains UPF0004	7	L to D
cce_3378	2		diurnal	**Two-component response regulator**	7	L to D
cce_1898	3	***patB***	circadian	**Transcriptional regulator (nitrogen fixation)**	17	D
cce_3594	4	***sigD***	circadian	**RNA polymerase sigma factor 2**	7	L to D
cce_0579	5	*fdxB*	circadian	Ferredoxin III	18	D
cce_4627	6	***tktA***	diurnal	**Transketolase**	1	D to L
cce_3149	7		circadian	unknown	1	D to L
cce_4205	8		circadian	hypothetical protein; contains a GCN5-related N-acetyltransferase domain	4	D
cce_3446	9		circadian	unknown	10	*
cce_3607	10		circadian	putative D-xylulose 5-phosphate/D-fructose 6-phosphate phosphoketolase	7	L to D
cce_3564	11		diurnal	unknown	3	L
cce_1844	12		circadian	unknown; contains an EF-Hand type domain	17	D
cce_1629	13	*glgP1*	circadian	Glycogen phosphorylase	10	*
cce_0043	14	*gmhA*	diurnal	Phosphoheptose isomerase	7	L to D
cce_2449	15		diurnal	unknown; contains a glycoside hydrolase, family 57 domain	1	D to L
cce_3617	16	*leuB*		3-isopropylmalate dehydrogenase	1	D to L
cce_0298	17	***rpaA***	diurnal	**Two-component response regulator (circadian rhythm)**	7	L to D
cce_1749	18		circadian	hypothetical protein; contains a conserved TM helix domain	7	L to D
cce_0072	19		diurnal	UPF YGGT-containing protein	3	L
cce_4510	20	*shc*	diurnal	Squalene-hopene-cyclase	7	L to D
cce_2625	21	*psbU*	diurnal	photosystem II 12 kD extrinsic protein	3	L
cce_2535	22	*opcA*	circadian	OxPPCycle protein	11	D
cce_2552	23		diurnal	unknown; contains amidinotransferase and CHP300 domains	1	D to L
cce_2500	24		circadian	hypothetical protein; contains a radical SAM domain	7	L to D
cce_1482	25		circadian	conserved hypothetical protein	11	D

Table 2 Prediction and function of coexpressed modules

Cluster	R (cyc)	Validation[a]	N	Enriched functions	Peak
1	0.78	0.98*	619	ribosomal proteins, chemotaxis, alanine and aspartate metbolism	D5-L1
2	0.33	0.66	90		L1
3	0.93	0.99*	173	diurnal, PSII, proteolysis and peptidolysis	L5-L9
4	0.85	−0.97	276	circadian, TCA cycle, reductive carboxylate cycle (CO_2 fixation), amino acid biosynthesis	L9-D5
5	0.67	0.98	14		D9
6	0.65	0.45	15		
7	0.57	0.97+	57	diurnal, photosystem I reaction center and PSI, aminosugars metabolism	L5-L9
11	0.55	0.99	70	circadian, cytochrome-c oxidase activity	L9-D1
12	0.60	−0.81	14	ribosome, porphyrin and chlorophyll metabolism	D5
13	0.46	0.96*	103	diurnal, phycobilisomes, photosynthesis antenna proteins, oxidative phosphorylation, RuBisCO, ATP synthase	L1
15	0.74	0.86	42	diurnal, peptidoglycan synthesis	L1-L5
17	0.43	0.74+	28	circadian, nitrogen fixation, nitrogenase	D1
18	0.67	−0.99	24	circadian, nitrogen fixation, nitrogenase, oxidoreductase activity	D1-D9
19	0.84	1.00*	11	circadian, nicotinate and nicotinamide metabolism, pentose-phosphate shunt	L9-D1

R (cyc), performance of cyclic model on cyclic data; Validation, performance of cyclic model on low oxygen data; N, number of genes in cluster; Enriched functions, statistically enriched functions in cluster ($p < 0.05$); Peak, time of peak expression in normal 12 h LD experiment. [a] Asterisks indicate statistical significance ($p < 0.02$) *versus* 100 random sets; plus indicates marginal significance ($p < 0.1$).

in light and this to be increased in the *sigD* mutant in *Synechocystis*.[35] *tktA* has a peak early in the day around L1 directly preceding induction of the photosynthetic machinery genes, consistent with these findings. The *opcA* gene is known to be essential for the functioning of the glucose-6-phosphate dehydrogenase (G6PDH) complex in the OPP pathway that plays an important role in glucose breakdown and generation of reducing power in several cyanobacteria.[37,38] This gene is located between genes associated with glycogen metabolism, all active in the dark and involved in breakdown of glycogen granules, and the nitrogenase gene cluster, which is powered by glycogen breakdown. Each of these examples from related organisms provides an important hypothesis about the functioning of the system in terms of transitions between system states in *Cyanothece* 51142.

Predictive model of transcriptomic dynamics in *Cyanothece*

One goal of a systems biology approach is to develop models that can provide accurate predictions of states of the system based on observation of a small number of inputs, for example environmental conditions or expression levels of regulators. We expanded on our previous analysis of the dynamics of *Cyanothece* 51142 under cyclic conditions[15] by developing a model that could predict the transcriptomic behavior of the system under novel conditions. Our prototype model used a

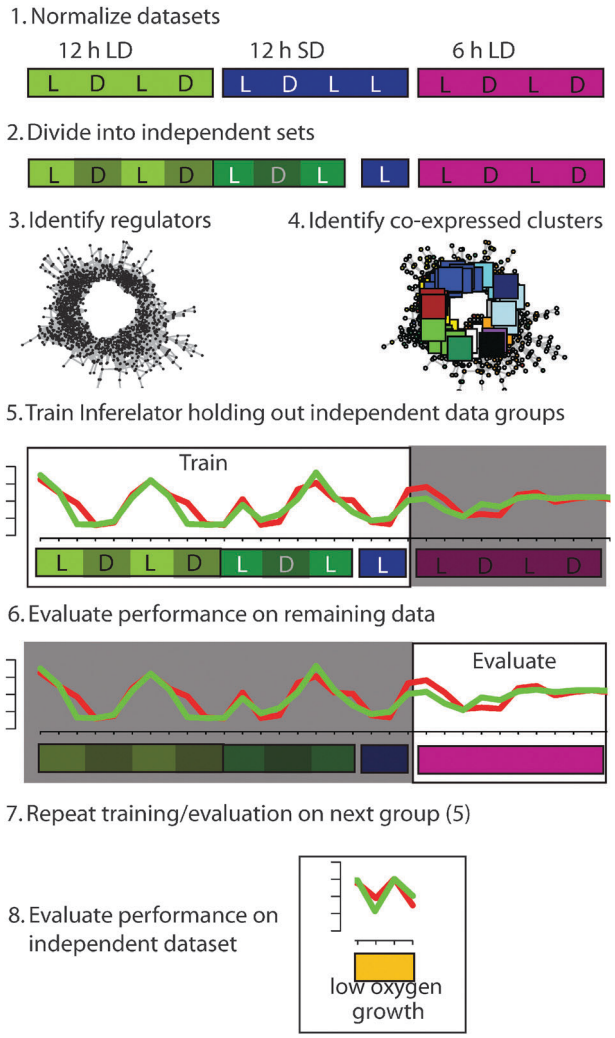


Fig. 2 Overview of predictive model construction and cross-validation procedure. (1) Transcriptomic data from three individual time course experiments, 12 h light/dark (LD), 12 h light/dark/continuous light (SD), and 6 h short day LD, were normalized and combined as described in the text. (2) To ensure reasonable cross-validation, redundant time points from the 12 h SD experiment were put together with identical time points in the 12 h LD experiment to establish independent sets. (3) Regulators were identified from annotations and topological analysis (see text). (4) Co-expressed clusters were identified from data to reduce the number of targets for inference. (5) A model is trained by holding out one independent set and training on the remaining data. (6) The resulting model is evaluated by predicting the behavior of the held-out set and comparing with observed behavior. (7) This process is then repeated for the other independent sets identified from step 2 to evaluate performance of the model in predicting new behavior. (8) Finally, an independent data set (growth under low oxygen, full-light conditions) is used to validate the model.

previously published multivariate regression method, the Inferelator,[30,39] to infer a network that relates the expression of minimal sets of regulators to the expression of target co-expressed clusters as ODEs.

Initially, we chose a set of potential regulators for model development that are annotated as transcription factors (TFs), sigma factors or circadian *kai* genes (see Table S2, ESI†). We defined the targets of the inference process to be co-expressed clusters of genes using various clustering methods and found the clustering approach and number of clusters that provides the best model performance (see Table S3, ESI†). Performance was assessed using a conservative cross-validation approach in which groups of similar conditions were treated independently for training and testing the model (Fig. 2), and are presented as the mean correlation between observed and predicted expression values per gene. The groups of datasets used and the maximum correlation between them and any of the other datasets are listed in Table S4.† This table shows that the groups are relatively independent of each other, with the maximum correlation between conditions in any two groups being no more than 0.6.

We found that our initial model, which includes 30 clusters as inference targets, provided very good performance as evaluated by cross-validation. The gene-normalized correlation for the model was found to be 0.62 indicating that the majority of genes were included in clusters whose behavior could be accurately predicted (Table 2). Functional enrichment of cluster membership show that these clusters represent previously observed functional groups (Table 2) that largely recapitulate our previous observations using network inference.[15] Since these functional groups are meant to provide a general idea of the functional processes that could be accurately predicted by our model, we have chosen to not impose a conservative multiple hypothesis correction, which leaves the majority of the functional processes listed passing such a filter. All targets/clusters were predicted with reasonable accuracy, and we show two examples of expression behavior for target clusters in Fig. 3.

Topological bottlenecks are predictive of system behavior

If the observed importance of topological bottlenecks in our inferred networks means that they are mediators of systems transitions, then expression profiles of bottlenecks should be predictive of target expression in our model. Accordingly, we examined how much predictive power could be attained using just the set of bottleneck genes as potential regulators. We used a set of bottlenecks (160 genes with the top 10% of betweenness values from the CLR-based network) as regulators to develop a model as described. This approach gave an overall performance of 0.70, better than that of the original model using only TFs (Table 3). This result indicates that bottlenecks are as effective at predicting the dynamics of the system as are the knowledge-based set of TFs, but it was unclear if the two sets were redundant or complementary in their ability to predict system behavior. We therefore combined the two sets of potential regulators (transcription/kai/sigma factors and topological bottlenecks) and found that the performance of the model increased to a correlation of 0.75, better than each set individually. Further, as a control, we replaced bottlenecks with sets of genes drawn at random from genes with low betweenness to use as regulators in addition to the transcription factors. Compared to the use of bottlenecks only as regulators, random sets of low-betweenness genes decreased the performance of the model significantly with the mean performance of 25 such models being 0.37. Combining these random sets with the TF set did not improve

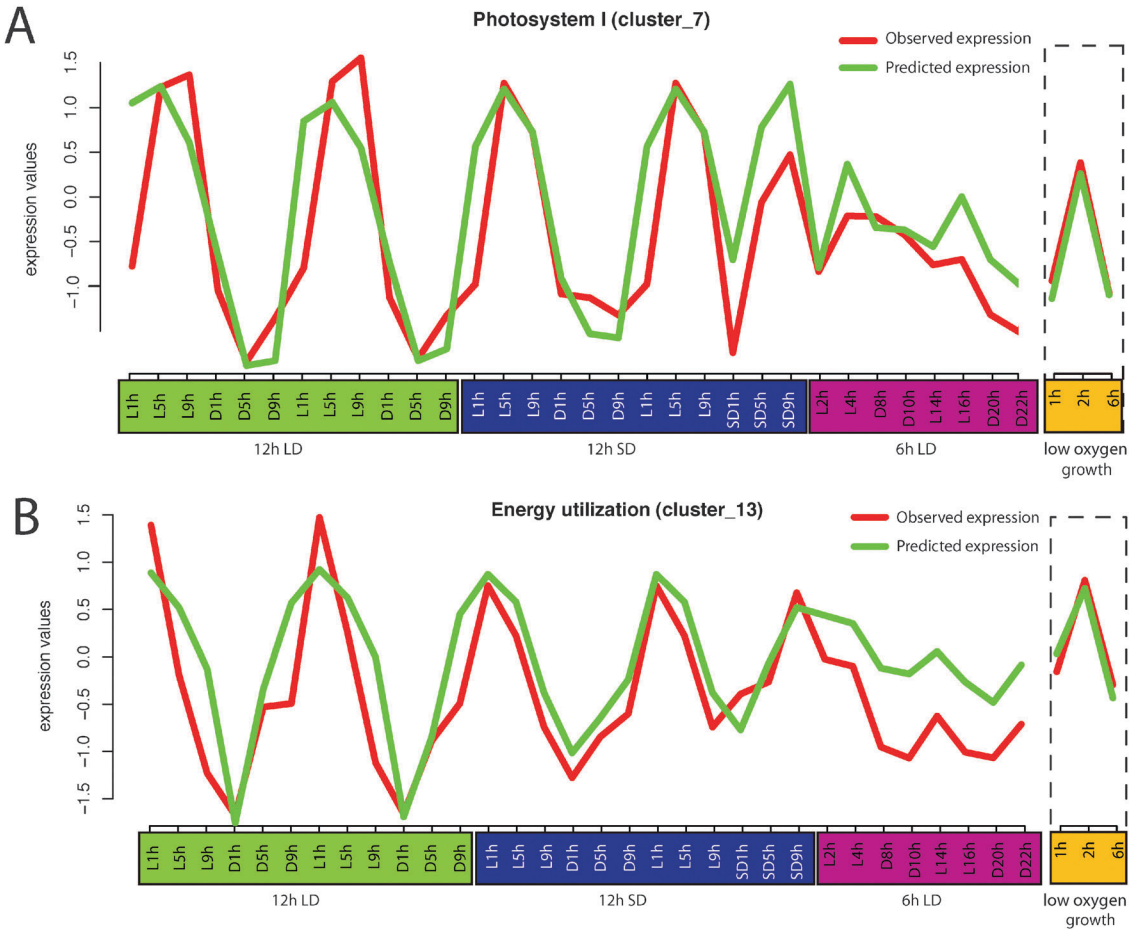


Fig. 3 Predicting the behavior of functional groups in *Cyanothece* 51142 over a range of conditions. We used the developed transcriptomic model to predict the behavior of all co-expressed clusters in *Cyanothece* 51142 using the cross-validation approach described (see Fig. 2 and text). The predicted (green) and observed (red) expression behavior is shown over the conditions used in this study (X axis). The colored bars represent the independent sets used for cross-validation (see Fig. 2). The dashed boxes show the performance of the model on the low oxygen, full-light experiment. (A). Expression of cluster_7 that contains most of the photosystem I complex and associated genes. These results show that the model can predict the transcriptional behavior of some targets with very high accuracy across a range of conditions not used for training. (B) Expression of cluster_13 that contains energy processing complexes, ATPase, RuBisCO, and Co-A biosynthesis.

Table 3 Topological bottlenecks predict system behavior as well as transcription factors

	All datasets		Validation dataset	
Dataset	Correlation	SD	Correlation	SD
TFs only	0.62		0.32	
Bottlenecks only	0.70		0.80	
TFs + bottlenecks	0.75		0.60	
TFs + random sets	0.62	0.03	0.11	0.38
Random sets	0.37	0.12	0.20	0.28

Correlation, mean correlation between predicted and observed expression levels per gene; SD, standard deviation.

performance at all. To examine whether these results were biased by consideration of cyclic data, we examined the impact of the same combinations on model performance for the low oxygen validation dataset. These results are more variable due to the lower number of conditions considered but strongly suggest that the bottlenecks play an important role in accurately predicting behavior of the validation data. This further supported the idea that bottlenecks play an important role in system function and that topological bottlenecks can be used to complement transcription and translation factors in building predictive models. It also shows that topological bottlenecks can predict system behavior well by themselves, in accordance with their elevated importance in the system.

Transcriptional regulatory structure of *Cyanothece* 51142

To determine the regulatory structure of *Cyanothece* 51142 during cyclic conditions, we used all the genes considered as regulators (TFs and bottlenecks) as targets of inference and used our modeling approach to determine a parsimonious set of relationships between these components. The high-confidence network combining this regulatory network with the regulator-target network described above is shown in Fig. S2, ESI.† We also used the CLR method, which determines regulatory networks using a mutual information approach, to infer a regulatory network using all transcriptional data. This network is very similar to that produced by our modeling approach, but it lacks the directionality of the regulatory relationships and

does not allow predictive cross-validation. Details about the construction of the CLR-based network are provided in SI along with a comparison between the two networks (Fig. S3, ESI†).

In particular, we focused on one portion of the regulatory network that involved three regulators with likely roles in important functional processes; *patB*, *rpaA*, and *ntcA*. Both the regression model and CLR-based regulatory networks predict *patB* to inhibit *rpaA* and *ntcA* (Fig. S3, ESI†). The regulator *ntcA* is thought to play a central role in regulation of heterocyst formation in *Anabaena* in response to nitrogen starvation.[40] 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,[41] 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.[36,42] In Fig. S2,† it can be seen that *ntcA* regulates cluster_17, the cluster that contains the *patB* gene. Therefore, it appears that there may be a feedback loop between *patB* and *ntcA*, which seem to play opposing roles in *Cyanothece* 51142.

Inferred regulatory influences accurately predict expression of nitrogenase and RuBisCO

Several clusters in our global model represent important complexes, including the nitrogenase (*nifHDK*) and RuBisCO (*rbcLS*). We show the inferred regulatory structure of both these complexes and the expression patterns of the cluster and the inferred regulatory influences in Fig. 4. The model predicts the expression of the core nitrogenase genes with a good correlation of predicted to observed expression of 0.67. The primary regulator that is predicted to influence the expression of the nitrogenase complex is PatB. The PatB TF is known to regulate transcription of nitrogenase in heterocysts in *Anabaena* sp. strain PCC 7120,[43] is co-conserved with nitrogenase across a number of cyanobacterial species,[41] and is a likely candidate as a regulator of nitrogenase in *Cyanothece* 51142. The nitrogenase activity is shown in Fig. 4A over the normal 12 h LD period, indicating that the gene expression patterns correlate well with activity for this complex, a well-established observation.[44]

The RuBisCO complex is formed by the *rbcS* and *rbcL* gene products and plays a central role in carbon fixation, which is closely linked to photosynthetic processes. The predicted regulatory influences on RuBisCO are shown in Fig. 4B and include an uncharacterized two-component regulator (cce_0678) that bears a strong resemblance to cce_0298 (*rpaA*). Both genes are similar to Ycf27 and Ycf29, chloroplast proteins that are found in all major plant and algal lineages and that encode similar transcription factors with a HTH DNA binding motif.[45–47] The exact function of these proteins is not known, but the parallelism of cce_0678 for photosynthesis (Fig. 4 and Fig. S3, ESI†) 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.

The levels of cyanophycin activity over the normal 12 h LD experiment are shown in Fig. 4B and correlate well with the gene expression. We also assessed the rate of CO_2 uptake in *Cyanothece* 51142 under 12 h LD conditions to correlate the uptake capacity for CO_2 with the RuBisCO gene expression. Data from this experiment revealed a peak expression of RuBisCO genes early in the light period and a maximum in CO_2 uptake late during the light period. This suggests that RuBisCO expression is anticipated by an increase in cellular CO_2 availability, and that its predicted regulators (including *cce_0678*) might be involved for this important process. However, further investigation is necessary to determine if RuBisCO expression is truly affected by CO_2 levels. These observations show that our approach to characterization of regulatory influences from high-throughput transcriptional data provides useful information about the function of complexes important in metabolism.

Model validation on non-cyclic expression data

The best evaluation of a predictive model is to apply it to data that has not been used in model training and is qualitatively different than that used to train the model. Accordingly, we examined the ability of the model to predict transcriptional behavior under low oxygen growth conditions that do not include LD transitions. *Cyanothece* 51142 was grown in the light without oxygen for 6 h and samples taken for transcriptomics at 1, 2, and 6 h (see Methods). Though portions of the response to low oxygen growth may be similar to that during the low oxygen conditions in the dark, the responses are substantially different because of the differences in growth conditions. The maximum correlation between the low oxygen conditions and any other training condition was 0.32, showing that the similarity between these conditions and any of the other training conditions is quite low (Table S4, ESI†). We evaluated the performance of the model trained on the cyclic time course data applied to the low oxygen time course data and found that the predictive performance was good (0.60 correlation observed *versus* predicted expression per gene) and that the behavior of many clusters could be very accurately predicted (see Table 2). Because there are a limited number of conditions in the validation set we were concerned that some of these results could be coincidental. Thus, we examined this possibility by randomly resorting gene labels for the validation set 100 times and calculating the *p*-values for the performance of the model on each. Significance ($p < 0.02$) is indicated in Table 2 and shows that the performance of two of the clusters with high performance on the validation data (clusters 5 and 11) did not pass our significance test, whereas the other highly predicted clusters did. The three clusters that are poorly predicted under low oxygen conditions (clusters 4, 12, and 18) seem to represent functions that are regulated very differently under light/dark, oxidative conditions *vs.* continuous light, low O_2 conditions (*e.g.* nitrogenase, CO_2 fixation), which may explain why the model fails to accurately capture their dynamics. Table 4 summarizes the overall performance of four models constructed from portions of the data, then

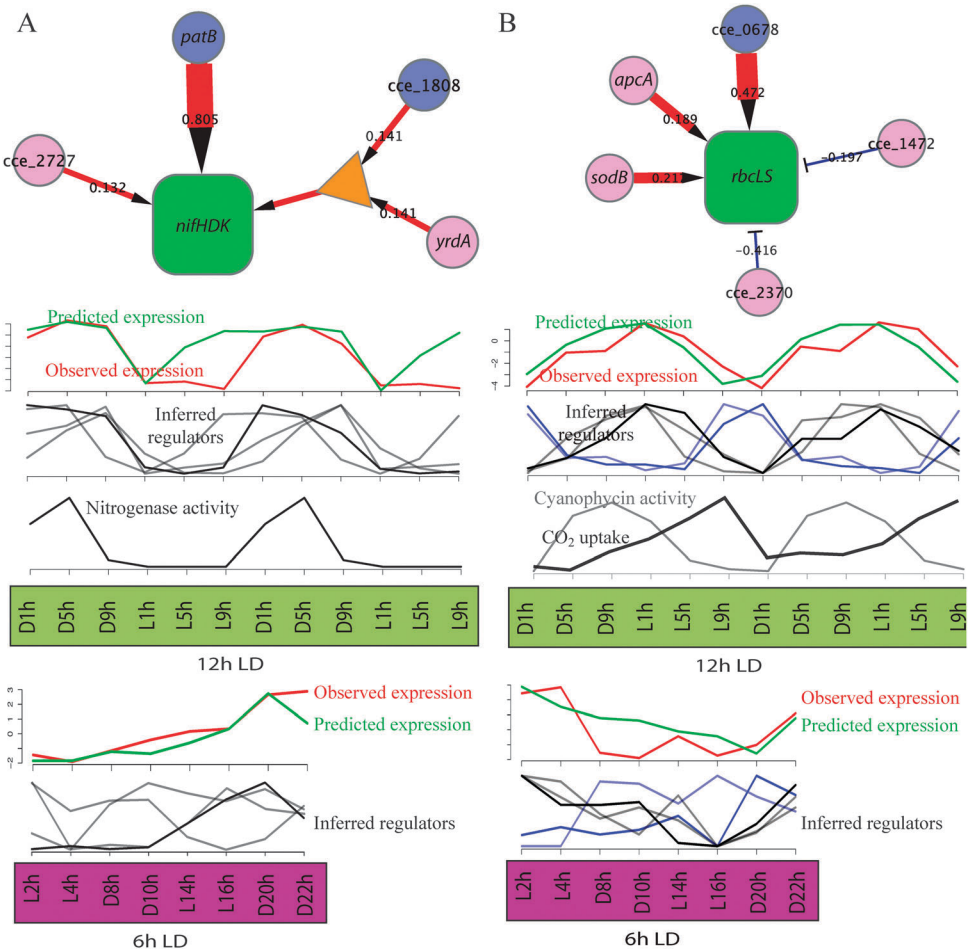


Fig. 4 Accurate prediction of nitrogenase and RuBisCO transcription. The inferred regulatory influences on the (A) core nitrogenase genes (*nif DHK*), and (B) RuBisCO complex (*rbcLS*) are shown with red arrows indicating positive influence and blue lines indicating negative influence. The orange triangle represents the influence is a combination of the two expression patterns. Blue nodes are genes that are transcriptional regulators, pink nodes are topological bottlenecks. The middle panels show expression patterns during the 12 h LD experiment: the predicted (green) and observed (red) expression for each cluster; the expression patterns of the inferred regulatory influences with black indicating the strongest positive influence (*patB* and *cce_0678* for nitrogenase and RuBisCO, respectively), and blue lines indicating negative influences; and levels of nitrogenase activity, CO_2 uptake and cyanophycin accumulation. The bottom panels show the predicted and observed expression patterns and regulator patterns for the 6 h LD experiment.

applied to the portion excluded from the model to independently validate the model. These results show that the models validate well with all independent data sets, but that the low oxygen data set has the lowest overall performance. This independent evaluation shows that the inferred model is consistent for most genes considered, but also highlights modules that require further experimental characterization.

Table 4 Performance on different independent validation datasets

Dataset	Training N	Testing N[a]	Correlation
Cyclic 12 h LD	14	21	0.78
Continuous Light	32	3	0.82
Short 6 h LD	27	8	0.80
Non cyclic low oxygen	32	3	0.60

Training N, number of conditions used to construct model; Testing N, number of conditions used to validate the model; Correlation, mean correlation between predicted and observed expression levels per gene.
[a] For testing groups see Table S4, ESI.†

Conclusions

In this study we have presented a predictive model of cyclic transcriptional processes in *Cyanothece* 51142, and show that this model can accurately predict the behavior of co-expressed clusters and important functional complexes under conditions not included in the training data. Additionally, we have extended our network analyses of the transcription of *Cyanothece* 51142 to highlight the importance of topological bottlenecks to the overall functioning of the system. Importantly, we show that topological bottlenecks are as good at predicting the behavior of the system as traditional regulators defined by gene annotation. Our results represent the first global predictive model of transcriptional behavior in a cyanobacterium.

The model we present can be queried in different ways to provide hypotheses pertinent to the functioning of the system as a whole, which can be validated experimentally. One kind of hypothesis is presented in this study, the predicted regulatory connections between functional components

(co-expressed clusters and other regulators) and can be validated by experiments that eliminate the activity of the predicted regulator and examine its effect on the expression of the predicted target. Prediction of the behavior of the system under novel conditions (as demonstrated in this study) is also possible, and the expression of a small number of regulators can predict the global behavior of the system. We are currently pursuing these avenues in *Cyanothece* 51142 and in the closely related and genetically tractable, *Cyanothece* sp. PCC 7822.

We have shown that topological analysis of association networks is promising for identification of true bottlenecks that mediate transitions between system states, identifying genes that are apparently more important to the system. These predictions summarize a large amount of information in the system, and thus represent a starting point for further investigation. They are based on analysis of high-throughput data, and therefore are unlikely by themselves to provide mechanistic insight into function. Examining the functions of connected genes in the network, temporally upstream and downstream, will shed light on the general function of the bottlenecks in the system. However, experimental investigation is needed to validate and further investigate these predictions.

The results we present in this study show that our modeling approach is very useful for understanding the regulation and dynamics of the transcriptomics of functional processes in a highly cyclic system. In some cases, the expression of genes directly reflects their function (for example the nitrogenase complex), however, this will not be true for all (or even most) cases. Therefore, integration of other data types, high-throughput proteomics and metabolomics for example, should provide the basis for a more complete model that can accurately predict more functional processes in the system.

Modules defined from the *Cyanothece* diurnal cycling transcriptomics data represent system states in which genes important for particular functions have peaks. The system requires regulatory and metabolic transitions to activate the appropriate system states in response to appropriate environmental signals.[48] This allows *Cyanothece* to be flexible in response to variations in photocycles and availability of nutrients. These transitions are mediated by transcriptional regulators, environmental sensors and proteins with other functions; *e.g.*, ion channels. The essential components of the system can then be thought of as the set of functional modules that actually do the work, and the mediators that join them together and regulate their activity. These 'mediators' of system transitions act as effectors that must be active under both the condition of origination and the 'target' condition. Mediators represent decision points where the system may choose to take a number of different courses based on the input signals, either environmental signals (light or dark) or inputs from the originating module. Our predictive model captures many of these elements, allowing accurate and robust prediction of transitions between system states.

Methods

Data sources

We used transcriptomic data from studies of *Cyanothece* 51142: 12 time points from a 12 h LD experiment sampled every 4 h over 48 h;[5] 12 time points from a 12 h dark/light/continuous light (subjective dark; SD) experiment sampled every 4 h over 48 h;[16] and 12 time points from a 6 h LD experiment sampled every 2 or 4 h over 24 h.[15] Datasets from each experiment were normalized using the standard Agilent array protocol, as described in the respective publications, and expressed as fold-change values from the mean expression value for each gene. The combined dataset was filtered to include only genes with fold-change greater than 2.5 in at least one condition. For display purposes the combined dataset was quantile normalized. This analysis resulted in 1595 genes with significantly changing expression profiles considered in our modeling efforts. The original microarray data for the 12 h LD and 12 h SD experiments are available through the European Bioinformatics Institute ArrayExpress (http://www.ebi.ac.uk/aerep/) database accession numbers E-TABM-337, and E-TABM-386. The microarray data for the 6 h LD and low oxygen experiments are currently being deposited in the ArrayExpress database.

Low oxygen growth conditions were previously described[49] and microarray data described separately.[15] Briefly, cells were grown under 12 h LD conditions with oxygen until time 0, when cells were bubbled with 99.9% N_2 and 0.1% CO_2, giving low-O_2 conditions. Cells were harvested for RNA preparation and microarray hybridization at 1, 2, and 6 h growth in low oxygen under full light.

Determination of homology

We used the program InParanoid[50] to determine protein homologs between *Cyanothece* 51142 and *Anabaena* sp. PCC 7120 (NC_003272), *Arabidopsis thaliana*,[51] *Escherichia coli* (NC_000913) and *Synechocystis*sp. PCC 6803 (NC_000911). For this study no distinction was made between orthologs and paralogs. Homologs were determined with InParanoid using a bit score threshold of 40, and considering protein pairs where the alignment covered more than 50% of each sequence.

Cyclic wreath construction

With the filtered set of expression profiles described above we calculated the Pearson correlation coefficient between all pairs of genes. A wreath network was generated by applying a stringent threshold (0.91) to the positive correlation values between all genes, as previously described for relevance networks.[5] Using standard graph layout algorithms (*e.g.* force-directed layout using the program "not" from the GraphViz suite; http://www.graphviz.org) the resulting networks are shaped like a wreath (see Fig. 1).

Network topology

We calculated network topology measures using the Python library NetworkX (http://networkx.lanl.gov/). Betweenness centrality is calculated as the percentage of times a node (gene) appears in the shortest path between all pairs of nodes. Node centrality is calculated as the number of neighbors of a particular gene (degree). Bottlenecks and hubs were defined as the top 20% of genes ranked by betweenness and node centrality, respectively, as previously described.[17,18,20]

Clustering

Hierarchical clustering was performed using hclust in the R statistical package and a range of distance methods and agglomeration methods (see ESI†). For Pearson and Spearman correlation a distance matrix was calculated as 1-R for all values.

Construction of predictive models

We used the Inferelator[30] version 1.1, as well as our own R code supporting the cross-validation approaches and other supporting utilities (available upon request) to develop predictive models based on transcriptomic profiles. We used sets of TFs (Table S2, ESI†) and topological bottlenecks (see Results) as potential regulators and sets of co-expressed genes identified using hierarchical clustering as the targets for inference. After assessing the performance of various clustering methods and hierarchical tree divisions resulting in different numbers of clusters, we found that a hierarchical clustering method using Euclidean distance between gene profiles and the 'mcquitty' agglomeration method and choosing 30 clusters provided the best performance (see Table S3, ESI†). Our training data was treated as three time courses in the Inferelator and we used a tau factor of 15 m for inference, as described previously for *Halobacterium*.[39] Though this tau may be short for our longer time intervals (2–4 h), models inferred with longer tau factors (30 and 60 m) produced poor results.

To provide a method for evaluating how well our models would generalize, that is, how well the models will be able to predict the behavior of the system under new conditions, we employed a cross-validation approach. For each evaluation of model performance, we trained four models independently: one trained on all data except the data gathered under standard 12 h LD cycle[5] and the data gathered from the first cycle of the continuous light experiment[16] as this was identical to the first experiment; one trained on all data except the continuous light time points; one trained on all data except the 6 h LD experimental data;[15] and one trained on all data except the three time points under low oxygen growth conditions for validation.[15] Each model was then used to predict the behavior of all targets, for those time points that were left out of the training set.

In models produced by the Inferelator the relation between the expression of a target (y) and the expression levels of regulators with non-null influences on y (X) is expressed as:

$$\tau \frac{dy}{dt} = -y + \sum \beta_i X_i \qquad (1)$$

Here, t is the time step used in model construction and b is the weight for relationship X on y as determined by L_1 shrinkage using least angle regression[52] in the Inferelator. Least angle regression selects a parsimonious set of predicted causal influences and learns their coefficients (b) from expression profiles.

We evaluated the ability of models to predict the average expression of each functional module given the expression levels of the regulators predicted to influence it. Assuming equilibrium conditions the derivative dy/dt is 0 and so eqn (1) can be represented simply as a linear weighted mean:

$$y = \sum \beta_j X_j \qquad (2)$$

We evaluated the performance of models by comparing the predicted expression levels of all targets with the observed expression levels using Pearson correlation. The overall performance of the model was calculated as the average performance of each target weighted by the number of genes represented by that target.

We first identified potential regulators in the *Cyanothece* genome based on their annotation in the genome[12] as a "regulator", we also included sigma factors and *kai* clock components in the analysis (Table S2, ESI†). We employed the Context Likelihood of Relatedness (CLR) method[32] applied to the expression profiles of the regulators over the four experiments listed above. The CLR method determines potential associations between regulators based on the mutual information metric between their profiles that also includes a filtering step that ranks a relationship between two genes based on its statistical significance relative to all the relationships determined for the two genes. The filtering step is designed to filter out indirect regulatory interactions and allowed confident inference of regulatory networks in *E. coli* previously.[32] CLR was performed using 10 bins for data discretization and a spline degree of 3.

Measurement of CO_2 uptake

Cyanothece 51142 cells grown under nitrogen fixing conditions in 12 h alternating LD were harvested by centrifugation and washed with air-saturated ASP2 medium without combined nitrogen. The cell pellet was resuspended in Hepes buffer (20 mM Hepes, 300 mM NaCl, pH 7.0) and adjusted to a chlorophyll concentration of 5 μg/mL. The CO_2 uptake measurements were performed using a WMA-4 CO_2 analyzer (PP Systems) at a flow rate of 1 L min^{-1}, a light intensity of 1000 μmol photons/m^2*s and a temperature of 30 °C. The CO_2 uptake was calculated as volume of fixed CO_2 in μmoles CO_2/mg Chl*h and is based on the assumption that 1 Mol of CO_2 equals 24 L at 30 °C.

Other experimental measures

Measurements of *Cyanothece* under 12 h light/dark cycles from previous publications were collated as follows: nitrogenase activity was taken from ref. 44 and 53; oxygen evolution and respiration were taken fromref. 44; carbohydrate levels were taken from ref. 54; and cyanophycin levels measured by Bradford assay, Western blot and electron microscopy were taken from ref. 53.

Functional enrichment

For functional enrichment analyses we used the automated pathway mapping from the Kyoto Encylopedia of Genes and Genomes (KEGG; ref. 55) and automated Gene Ontology assignments from InterPro domains[56] using the Bioverse annotation pipeline.[57]

We calculated functional enrichment (*e.g.* of identified clusters) by considering each functional label individually and calculating the chi-square test value between the representation of the function in the cluster or group of interest *versus* all the other genes in the network. Only genes

with a functional label were used in this analysis. A p-value of 0.05 or less was considered to be significant.

Determination of cyclic behavior

To more clearly determine the cyclic nature of genes in the *Cyanothece* dataset we used the online Haystack tool described in ref. 8 at http://haystack.cgrb.oregonstate.edu. We used the default significance criteria provided by Haystack (p value < 0.05) and used a correlation coefficient filter of 0.8 for evaluation of cyclic genes. Diurnal cyclic genes were identified as those genes that were cyclic in the 12 h LD experiment, but not in the 12 h SD or the 6 h LD experiments whereas circadian genes were cyclic in all three datasets. Although this is a conservative criterion for classifying genes with circadian rhythm it highlights genes that are under robust circadian control.

Acknowledgements

We would like to thank Jörg Toepel for his effort in generating some of the microarray data. This work is part of a Membrane Biology EMSL Scientific Grand Challenge project at the W.R. Wiley Environmental Molecular Sciences Laboratory, a national scientific user facility sponsored by U.S. Department of Energy's Office of Biological and Environmental Research (BER) program located at Pacific Northwest National Laboratory (PNNL). PNNL is operated for the U.S. Department of Energy by Battelle.

References

1 D. Bell-Pedersen, V. M. Cassone, D. J. Earnest, S. S. Golden, P. E. Hardin, T. L. Thomas and M. J. Zoran, *Nat. Rev. Genet.*, 2005, **6**, 544–556.
2 R. Aurora, Y. Hihara, A. K. Singh and H. B. Pakrasi, *OMICS*, 2007, **11**, 166–185.
3 R. T. Gill, E. Katsoulakis, W. Schmitt, G. Taroncher-Oldenburg, J. Misra and G. Stephanopoulos, *J. Bacteriol.*, 2002, **184**, 3671–3681.
4 M. A. Woelfle and C. H. Johnson, *J. Biol. Rhythms*, 2006, **21**, 419–431.
5 J. Stockel, E. A. Welsh, M. Liberton, R. Kunnvakkam, R. Aurora and H. B. Pakrasi, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 6156–6161.
6 C. E. Boothroyd, H. Wijnen, F. Naef, L. Saez and M. W. Young, *PLoS Genet.*, 2007, **3**, e54.
7 H. Wijnen, F. Naef, C. Boothroyd, A. Claridge-Chang and M. W. Young, *PLoS Genet.*, 2006, **2**, e39.
8 T. P. Michael, T. C. Mockler, G. Breton, C. McEntee, A. Byer, J. D. Trout, S. P. Hazen, R. Shen, H. D. Priest, C. M. Sullivan, S. A. Givan, M. Yanovsky, F. Hong, S. A. Kay and J. Chory, *PLoS Genet.*, 2008, **4**, e14.
9 M. A. Woelfle, Y. Xu, X. Qin and C. H. Johnson, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 18819–18824.
10 G. Dong and S. S. Golden, *Curr. Opin. Microbiol.*, 2008, **11**, 541–546.
11 S. R. Mackey and S. S. Golden, *Trends Microbiol.*, 2007, **15**, 381–388.
12 E. A. Welsh, M. Liberton, J. Stockel, T. Loh, T. 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, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 15094–15099.
13 D. G. Adams, *Curr. Opin. Microbiol.*, 2000, **3**, 618–624.
14 K. J. Reddy, J. B. Haskell, D. M. Sherman and L. A. Sherman, *J. Bacteriol.*, 1993, **175**, 1284–1292.
15 J. Toepel, J. McDermott, T. C. Summerfield and L. A. Sherman, *J. Phycol.*, 2009, **45**, 610–620.
16 J. Toepel, E. Welsh, T. C. Summerfield, H. B. Pakrasi and L. A. Sherman, *J. Bacteriol.*, 2008, **190**, 3904–3913.
17 H. Yu, P. M. Kim, E. Sprecher, V. Trifonov and M. Gerstein, *PLoS Comput. Biol.*, 2007, **3**, e59.
18 M. D. Dyer, T. M. Murali and B. W. Sobral, *PLoS Pathog.*, 2008, **4**, e32.
19 L. Yao and A. Rzhetsky, *Genome Res.*, 2008, **18**, 206–213.
20 J. E. McDermott, R. C. Taylor, H. Yoon and F. Heffron, *J. Comput. Biol.*, 2009, **16**, 169–180.
21 H. Yoon, J. E. McDermott, S. Porwollik, M. McClelland and F. Heffron, *PLoS Pathog.*, 2009, **5**, e1000306.
22 D. L. Diamond, A. J. Syder, J. M. Jacobs, C. M. Sorensen, K. A. Walters, S. C. Proll, J. E. McDermott, M. A. Gritsenko, Q. Zhang, R. Zhao, T. O. Metz, D. G. Camp, 2nd, K. M. Waters, R. D. Smith, C. M. Rice and M. G. Katze, *PLoS Pathog.*, 2010, **6**, e1000719.
23 J. Cerveny and L. Nedbal, *J. Biol. Rhythms*, 2009, **24**, 295–303.
24 M. R. Roussel, D. Gonze and A. Goldbeter, *J. Theor. Biol.*, 2000, **205**, 321–340.
25 F. L. Hellweger, *Ecol. Modell.*, 2010, **221**, 1620–1629.
26 H. Knoop, Y. Zilliges, W. Lockau and R. Steuer, *Plant Physiol.*, 2010, **154**, 410–422.
27 A. K. Singh, T. Elvitigala, J. C. Cameron, B. K. Ghosh, M. Bhattacharyya-Pakrasi and H. B. Pakrasi, *BMC Systems Biology*, 2010, **4**, 105.
28 Z. Su, F. Mao, P. Dam, H. Wu, V. Olman, I. T. Paulsen, B. Palenik and Y. Xu, *Nucleic Acids Res.*, 2006, **34**, 1050–1065.
29 S. Okamoto, Y. Yamanishi, S. Ehira, S. Kawashima, K. Tonomura and M. Kanehisa, *Proteomics*, 2007, **7**, 900–909.
30 R. Bonneau, D. J. Reiss, P. Shannon, M. Facciotti, L. Hood, N. S. Baliga and V. Thorsson, *GenomeBiology*, 2006, **7**, R36.
31 J. E. McDermott, M. Costa, D. Janszen, M. Singhal and S. C. Tilton, *Dis. Markers*, 2010, **28**, 253–266.
32 J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins and T. S. Gardner, *PLoS Biol.*, 2007, **5**, e8.
33 C. Caretta-Cartozo, P. De Los Rios, F. Piazza and P. Lio, *PLoS Comput. Biol.*, 2007, **3**, e103.
34 J. Liang, L. Scappino and R. Haselkorn, *J. Bacteriol.*, 1993, **175**, 1697–1704.
35 T. C. Summerfield and L. A. Sherman, *J. Bacteriol.*, 2007, **189**, 7829–7840.
36 N. Takai, M. Nakajima, T. Oyama, R. Kito, C. Sugita, M. Sugita, T. Kondo and H. Iwasaki, *Proc. Natl. Acad. Sci. U. S. A.*, 2006, **103**, 12109–12114.
37 M. L. Summers, J. G. Wallis, E. L. Campbell and J. C. Meeks, *Journal of Bacteriology*, 1995, **177**, 6184–6194.
38 A. K. Singh, H. Li, L. Bono and L. A. Sherman, *Photosynth. Res.*, 2005, **84**, 65–70.
39 R. Bonneau, M. T. Facciotti, D. J. Reiss, A. K. Schmid, M. Pan, A. Kaur, V. Thorsson, P. Shannon, M. H. Johnson, J. C. Bare, W. Longabaugh, M. Vuthoori, K. Whitehead, A. Madar, L. Suzuki, T. Mori, D. E. Chang, J. Diruggiero, C. H. Johnson, L. Hood and N. S. Baliga, *Cell*, 2007, **131**, 1354–1365.
40 J. W. Golden and H. S. Yoon, *Curr. Opin. Microbiol.*, 2003, **6**, 557–563.
41 K. Stucken, U. John, A. Cembella, A. A. Murillo, K. Soto-Liebe, J. J. Fuentes-Valdes, M. Friedel, A. M. Plominsky, M. Vasquez and G. Glockner, *PLoS One*, 2010, **5**, e9235.
42 M. K. Ashby and C. W. Mullineaux, *FEMS Microbiol. Lett.*, 1999, **181**, 253–260.
43 K. M. Jones, W. J. Buikema and R. Haselkorn, *J. Bacteriol.*, 2003, **185**, 2306–2314.
44 M. S. Colon-Lopez, D. M. Sherman and L. A. Sherman, *J. Bacteriol.*, 1997, **179**, 4319–4327.
45 S. Puthiyaveetil and J. F. Allen, *Proc. Biol. Sci.*, 2009, **276**, 2133–2145.
46 S. Puthiyaveetil, T. A. Kavanagh, P. Cain, J. A. Sullivan, C. A. Newell, J. C. Gray, C. Robinson, M. van der Giezen, M. B. Rogers and J. F. Allen, *Proc. Natl. Acad. Sci. U. S. A.*, 2008, **105**, 10061–10066.
47 M. K. Ashby, J. Houmard and C. W. Mullineaux, *FEMS Microbiol. Lett.*, 2002, **214**, 25–30.
48 A. Mitchell, G. H. Romano, B. Groisman, A. Yona, E. Dekel, M. Kupiec, O. Dahan and Y. Pilpel, *Nature*, 2009, **460**, 220–224.

49 T. C. Summerfield, J. Toepel and L. A. Sherman, *Biochemistry Rapid Reports*, 2008, **74**, 12939–12941.
50 M. Remm, C. E. Storm and E. L. Sonnhammer, *J. Mol. Biol.*, 2001, **314**, 1041–1052.
51 D. Swarbreck, C. Wilks, P. Lamesch, T. Z. Berardini, M. Garcia-Hernandez, H. Foerster, D. Li, T. Meyer, R. Muller, L. Ploetz, A. Radenbaugh, S. Singh, V. Swing, C. Tissier, P. Zhang and E. Huala, *Nucleic Acids Res.*, 2008, **36**, D1009–1014.
52 B. Efron, I. Johnstone, T. Hastie and R. Tibshirani, *Annals of Statistics*, 2003, **32**, 407–499.
53 H. Li, D. M. Sherman, S. Bao and L. A. Sherman, *Arch. Microbiol.*, 2001, **176**, 9–18.
54 M. A. Schneegurt, D. M. Sherman and L. A. Sherman, *Arch. Microbiol.*, 1997, **167**, 89–98.
55 M. Kanehisa, S. Goto, S. Kawashima and A. Nakaya, *Nucleic Acids Res.*, 2002, **30**, 42–46.
56 R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, E. Birney, M. Biswas, P. Bucher, L. Cerutti, F. Corpet, M. D. Croning, R. Durbin, L. Falquet, W. Fleischmann, J. Gouzy, H. Hermjakob, N. Hulo, I. Jonassen, D. Kahn, A. Kanapin, Y. Karavidopoulou, R. Lopez, B. Marx, N. J. Mulder, T. M. Oinn, M. Pagni, F. Servant, C. J. Sigrist and E. M. Zdobnov, *Bioinformatics*, 2000, **16**, 1145–1150.
57 J. McDermott, M. Guerquin, Z. Frazier, A. N. Chang and R. Samudrala, *Nucleic Acids Res.*, 2005, **33**, W324–325.

