1050–1065 *Nucleic Acids Research, 2006, Vol. 34, No. 3*
doi:10.1093/nar/gkj496

Computational inference and experimental validation of the nitrogen assimilation regulatory network in cyanobacterium *Synechococcus* sp. WH 8102

Zhengchang Su[1,2], Fenglou Mao[1], Phuongan Dam[1,2], Hongwei Wu[1,2], Victor Olman[1], Ian T. Paulsen[3], Brian Palenik[4] and Ying Xu[1,2,*]

[1]Department of Biochemistry and Molecular Biology, University of Georgia, Athens, GA 30602, USA, [2]Computational Biology Institute, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA, [3]The Institute of Genome Research, Rockville, MD 20850, USA and [4]Scripps Institution of Oceanography, University of California at San Diego, San Diego, CA 92093, USA

Received October 6, 2005; Revised December 20, 2005; Accepted January 23, 2006

ABSTRACT

Deciphering the regulatory networks encoded in the genome of an organism represents one of the most interesting and challenging tasks in the post-genome sequencing era. As an example of this problem, we have predicted a detailed model for the nitrogen assimilation network in cyanobacterium *Synechococcus* sp. WH 8102 (WH8102) using a computational protocol based on comparative genomics analysis and mining experimental data from related organisms that are relatively well studied. This computational model is in excellent agreement with the microarray gene expression data collected under ammonium-rich versus nitrate-rich growth conditions, suggesting that our computational protocol is capable of predicting biological pathways/networks with high accuracy. We then refined the computational model using the microarray data, and proposed a new model for the nitrogen assimilation network in WH8102. An intriguing discovery from this study is that nitrogen assimilation affects the expression of many genes involved in photosynthesis, suggesting a tight coordination between nitrogen assimilation and photosynthesis processes. Moreover, for some of these genes, this coordination is probably mediated by NtcA through the canonical NtcA promoters in their regulatory regions.

INTRODUCTION

From a systems biology's point of view, cellular functions are carried out through the interactions of molecules in a cell, with genes and their products being the major players (1). Such complex interaction networks in a cell are generally made of functional modules with different levels of complexity. At the bottom of this functional hierarchy are the basic biochemical pathways, and the coordinated interactions among these pathways are at the next level for carrying out more complex biological functions (1,2). The availability of the rapidly expanding pool of sequenced genomes and other high-throughput biological data, such as microarray gene expression data and proteomic data has made it possible to computationally infer the complex biological networks in a systematic manner (3).

Cyanobacterial species/strains in the genera of *Synechococcus* and *Prochlorococcus,* including *Synechococcus* sp. WH 8102 (WH8102), *Prochlorococcus* MED4 (MED4), *Prochlorococcus* MIT 9313 (MIT9313) and *Prochlorococcus* CCMP1375 (CCMP1375), are among the major oxygenic phototrophic strains living in a wide range of oceanographic areas, and they together contribute to a great portion of global CO_2 fixation (4,5). Therefore, a better understanding of these organisms may help solve some of the challenging global environmental problems. To facilitate systematic studies of these species/strains, the genomes of these organisms have been sequenced (6–8). As of today, our understanding of these organisms remains largely at the level of conventional genome annotation and physiological studies.

To facilitate the characterization of microbial organisms at a systems level, we have been developing a computational

*To whom correspondence should be addressed at Department of Biochemistry and Molecular Biology, A110 Life Sciences Building, 120 Green Street, University of Georgia, Athens, GA, 30602. Tel: +1 706 542 9779; Fax: +1 706 542 9751; Email: xyn@bmb.uga.edu

protocol for the inference of metabolic and regulatory networks in bacteria (9). This protocol consists of a number of inference steps and associated computational tools for prediction of operon structures, pathway mapping across genomes, prediction of *cis*-regulatory binding sites, prediction of functional association among proteins and information integration for building the network models (9). Using this computational capability, we have previously predicted the wiring diagrams for the phosphorus assimilation network (9) and carbon fixation network (10) in WH8102. As a continued effort to characterize the major regulatory networks in this organism, we present our work on the prediction of the nitrogen assimilation network in WH8102, and its verification and refinement using microarray gene expression data.

Nitrogen is an important element for all forms of life, and is limiting in the oliogotrophic oceanographic areas where WH8102 exists (11). It is generally known that a number of genes involved in nitrogen assimilation in cyanobacteria are regulated by the global regulator NtcA, which belongs to the cAMP receptor protein (CRP) family (12). Thus, this is different from the situation in proteobacteria such as *Escherichia coli*, where nitrogen assimilation related genes are regulated by the two-component system NtrB–NtrC (13). We have recently conducted a computational analysis of the NtcA regulons in nine sequenced cyanobacterial genomes including WH8102, and have predicted possible members of these regulons (14). It should be noted that not all nitrogen assimilation related genes in cyanobacteria are directly regulated by NtcA (15). For instance, the nitrogen starvation-induced global response called chlorosis involves more proteins than those that are directly regulated by NtcA (16–18). The goal of the current work is to gain a general and better understanding of the nitrogen assimilation network in WH8102 beyond the NtcA regulon (14), using a combined approach of computational prediction and experimental validation. To the best of our knowledge, no experimental work has been previously reported on the nitrogen assimilation network in WH8102.

MATERIALS AND METHODS

Materials

The genome and protein sequences of 231 prokaryotes were retrieved from the NCBI website (ftp://ftp.ncbi.nih.gov) or from Oak Ridge National Laboratory's genome channel database (http://compbio.ornl.gov/channel/). The names of these genomes are listed at http://csbl.bmb.uga.edu/~zhx/pathways/nitrogen/. The pair-wise protein–protein interaction datasets for *Helicobacter pylori, E.coli* and *Saccharomyces cerevisiae* were retrieved from the DIP database (http://dip.doe-mbi.ucla.edu/).

Prediction of operons/transcription units

Multi-gene operons are predicted using our JPOP program (19,20) for each of the eleven cyanobacterial genomes, namely, *Gloeobacter violaceus*, *Nostoc* sp. PCC 7120, *Prochlorococcus marinus* CCMP1375, *P.marinus* MED4, *P.marinus* MIT 9313, *Synechococcus elongatus* PCC 6301, *Synechococcus* sp. WH 8102, *Synechococcus* sp. CC9605, *Synechococcus* sp. CC9902, *Synechocystis* sp. PCC 6803 and *Thermosynechococcus elongatus*. For genes arranged in tandem on the same strand and not predicted to be part of an operon by JPOP, we predict them to form an operon if their intergenic distances are shorter than 45 bp. Genes that are not covered by the above procedure are predicted to form each a single-gene transcription unit.

Network inference protocol

We follow a similar protocol to that described in (9) for the inference of the nitrogen assimilation network in WH8102. The overall protocol consists of four key steps: (i) construction of template network models for related organisms where the target network is relatively well studied; (ii) mapping predicted template networks to the target genome, and construction of an initial network model; (iii) expansion of the initial network model; and (iv) validation and refinement of the predicted model using experimental data. Since the publication of this computational protocol (9), a number of new features have been added as described below.

Mapping template networks onto the target genome and construction of the initial network model. For this study, we have built the template network models based on extensive literature searches, in four related organisms: *Nostoc* sp. PCC 7120 (PCC7120), *Synechocystis* sp. PCC 6803 (PCC6803), *Synechococcus* sp. PCC 7942 (PCC7942) and *Synechococcus* sp. WH 8103 (WH8103). Then each template model is mapped onto the target genome WH8102 using the P-MAP program (21), which produces a (partial) model for the target network. P-MAP maps a template network model onto a target genome by finding the orthologous relationships between the genes in the template genome and those in the target genome using both sequence similarity information and operon structure information (21). Our previous study has shown that using these two pieces of information improves the mapping accuracy compared to tools based solely on sequence similarity information such as COG (22) or the bidirectional best hit method (23). For a template network in an organism whose genome sequence is not available, we simply map the genes in the template onto the target genome using the BLASTP program by using the best hit with an *E*-value $<10^{-20}$.

An initial network model is built by doing a union operation over the four mapped network models from the four template networks. If different functional roles are assigned to a gene in the target genome by different mappings, we resolve the problem by using the gene assignment with the highest confidence score (lowest *E*-value from BLASTP search) or through phylogenetic tree analysis using the TREE-PUZZZLE program (24) to select the one with the closest evolutionary distance.

Expansion of the initial network model. Our method then uses an 'expansion' step to recruit additional genes into the initial network model through application of a simple 'guilty by association' rule. This is done through prediction of physical interactions or functional associations between proteins already in the network model and those not in the model. The basic idea is that if protein A is already in the initial model but B is not, we will consider adding B to the model if A and B are predicted to either physically interact or be functionally associated. We will either predict the detailed interaction relationship between B and the rest of the network model or simply indicate that B is a part of the model, depending on the information available. Currently, we predict such physical interactions or functional associations using the following methods: (i) prediction of operons: if protein A of operon O is in the initial network model, then all other members of operon O are predicted to be part of the network; (ii) prediction of physical interactions between proteins; (iii) prediction of co-regulated transcription units (regulon); and (iv) prediction of functional association of proteins through phylogenetic profile analysis. We apply this expansion operation to each protein in the initial network model that does not conflict with our general knowledge of the network that is being inferred.

Prediction of physical interactions between proteins

(i) *Prediction through orthology mapping*: there are a number of public databases of protein–protein interactions derived from large-scale two-hybrid experiments (25), mass spectrometry (26) as well as from individual experiments over the years (27). The most comprehensive such datasets are for *S.cerevisiae* (25,26) and *H.pylori* (28). Our orthology mapping method predicts that two proteins interact if their orthologues are known to interact in *S.cerevisiae* or *H.pylori*, based on information stored in the DIP database (27), which contains 15 410 and 1420 pair-wise interactions for *S.cerevisiae* and *H.pylori*, respectively (release of July 4, 2004; http://dip.doe-mbi.ucla.edu/). A confidence score $S(i,j)$ is assigned to a predicted interaction between proteins i and j as follows.

$$S(i,j) = -\frac{1}{2}(\log E(i,a) + \log E(j,b),$$

where $E(i,a)$ and $E(j,b)$ are the E-values of the proteins i and j aligned to the known interacting proteins a and b in the database, respectively, by BLASTP. If an E-value is 0, we assign 10^{-250} to it for practical purposes. If the proteins i and j are mapped to multiple known interacting partners in different species, then the highest $S(i,j)$ is used.

(ii) *Prediction through protein fusion analysis*: the basic idea of the protein fusion analysis is that if two proteins A and B are homologous to different segments (domains) of the same protein chain C encoded in another genome, then A and B are predicted to interact with each other (29,30). We used BLASTP to search all the open reading frames (ORFs) of WH8102 against the non-redundant (*nr*) protein database containing 1 539 396 sequences (release of October 2003, ftp://ftp.ncbi.nih.gov/blast/db/). Proteins i and j in WH8102 are predicted to physically interact with each other, if they are homologous (E-values <0.0001) to different domains of the same protein in *nr*. A confidence score $T(i,j)$ is assigned to each pair of predicted interacting proteins i and j as follows.

$$T(i,j) = -\frac{1}{2}[\log E(i,a) + \log E(j,b)],$$

where $E(i,a)$ and $E(j,b)$ are the best E-values of the proteins i and j aligned to the segments a and b respectively, of a protein in the *nr* database by BLASTP. If an E-value is 0, we assign 10^{-250} to it.

To test the quality of our predictions by these two methods, we predicted protein–protein interactions in *E.coli* K12 that has a relatively large number of known pair-wise interactions in the DIP database (http://dip.doe-mbi.ucla.edu/). We calculate a P-value for the predictions by each method at a confidence score cutoff s_c, based on the null hypothesis that proteins in a genome interact with one another randomly. Let N be the total number of possible interactions in a genome with n genes, so $N = n(n+1)/2$. Let K_P be the number of predicted interactions in the genome with a score above the cutoff s_c, K_E be the total number of the known interactions, and K_V be the number of experimentally verified interactions among the predicted at the cutoff s_c, Then the P-value is calculated using the following hypergeometric cumulative function

$$p(K_V \mid s_c) = 1 - \sum_{i=1}^{K_V} \frac{\binom{K_E}{i}\binom{N-K_E}{K_P-i}}{\binom{N}{K_P}}.$$

For *E.coli* K12, $n = 4311$, thus $N = 9\,294\,516$ and $K_E = 457$ (http://dip.doe-mbi.ucla.edu/). The hygecdf procedure in Matlab package (The MathWorks, Inc.) was used to compute the P-values.

Prediction of co-regulated transcription units (regulon)

We used an improved phylogenetic footprinting technique (31–33) to predict the NtcA promoters in WH8102 and other 10 sequenced cyanobacterial genomes. The detail of the algorithm is described elsewhere (14). Briefly, we first identified in the initial network model the genes that are likely to be up-regulated by NtcA, and then identified the orthologues of these genes in the other 10 cyanobacterial genomes. We pooled the entire upstream regions (if they are longer than 800 bases, then only the immediate upstream 800 bases were used) of these orthologues in each genome according to the predicted operons/transcription units. If an intergenic region is shorter than 100 bases, then 10 bases in its upstream coding region was included. Then, single or multiple co-occurring binding sites were searched in these pooled sequences based on the knowledge about the promoters. In the case of NtcA promoters, it is known that NtcA binds to a pseudo-palindromic consensus GTAN$_8$TAC motif, and that an *E.coli*, -10 consensus-like box in the form of TAN$_3$T/A downstream to the NtcA binding is also required for the genes up-regulated by NtcA (34). Hence, the two motifs were searched using the CUBIC program (35). Then sequence profiles were built for these two binding sites based on the identified sequences. Single or multiple profiles were used to scan all the intergenic sequences for additional binding sites using a scoring function, which combines the information of multiple putative binding sites (if multiple profiles were used) in the promoter region of a gene and the presence of similar binding sites in the promoter regions of its orthologous genes in other related genomes. A P-value cutoff 0.05 is used for the prediction of putative NtcA promoters (for the simplicity of discussion, an NtcA promoter in this paper is defined as an NtcA binding site plus a downstream -10 like box in contrast to a single pseudo-palindromic NtcA binding site).

Prediction of functional association of proteins through phylogenetic profile analysis

The phylogenetic profile analysis predicts that two proteins are functionally related if they have highly similar phylogenetic profiles (36). Using a set of m reference genomes $\{G_1,..., G_m\}$, a protein's phylogenetic profile is defined as a binary string $a_1, ..., a_m$ with $a_i = 1$ if the protein has a detectable homologue in genome G_i (defined by BLASTP using an E-value $<10^{-6}$ in both directions of search); $a_i = 0$ otherwise. We have used $m = 231$ sequenced prokaryotic genomes (the list of the genomes is available at http://csbl.bmb.uga.edu/~zhx/pathways/nitrogen/) to compute the phylogenetic profiles for all the proteins of WH8102. We then clustered the proteins based on the similarities of their phylogenetic profiles, using an algorithm described as follows. We define a fully connected graph G (V, E) with V being the set of vertices, each of which represents a protein encoded in the target genome, and E being the set of edges connecting the proteins. The weight of the edge between vertices v_i and v_j (representing proteins g_i and g_j) is the distance between their phylogenetic profiles p_i and p_j with length $L = 231$, defined as follows.

$$d(p_i,p_j) = \frac{d_H}{1+\alpha C},$$

$$C = -p\log p - (1-p)\log(1-p),$$

$$p = \frac{h}{L-d_H}, \pm$$

where d_H is the Hamming distance between p_i and p_j, α is a constant (we choose $\alpha = 2$ in this study), and h is the number of genomes that contain the orthologues of both g_i and g_j. We now outline an algorithm for identifying protein clusters with similar phylogenetic profiles.

We first build a minimum spanning tree T of $G(V, E)$, using Prim's algorithm (37). Each vertex v is assigned an integer number $k \in \{1,2,\ldots,n\}$, if v is the kth selected vertex by Prim's algorithm when building the minimum spanning tree, where n is the number of proteins encoded in the target genome. We call k Prim's mapping coordinate of v, or simply, mapping coordinate of v. We draw a 2D plot as follows. The horizontal axis represents the mapping coordinates of all the vertices of G. For coordinate k on the horizontal axis, its corresponding value on the vertical axis represents the weight of the edge between the kth selected vertex v_k and the vertex that 'recruits' the kth vertex into the minimum spanning tree. We call this weight the inclusion distance (d_I) of v_k. We have previously demonstrated (35) that a group of proteins form a cluster (based on the similarities of their phylogenetic profiles) if and only if these proteins form a 'valley' in this 2D plot. We use the same method of (35) to evaluate the statistical significance of each identified cluster. If a protein in the initial network model belongs to a cluster with a statistical significant value (P-value) less than a pre-selected threshold (0.05 is used in this study), we predict that all the other proteins in this cluster are also involved in the network.

Validation and refinement of the network model. To validate our constructed nitrogen assimilation network model, we compared the genes in our predicted model to the microarray gene expression data collected under ammonium-rich versus nitrate-rich growth conditions (see below). We used the following cumulative hypergeometric distribution function to evaluate the statistical significance of the predictions by each step/method in our protocol, based on the null hypothesis that genes are randomly affected by the experimental manipulation.

$$p(K,V) = 1 - \sum_{i=1}^{V} \frac{\binom{A}{i}\binom{n-A}{K-i}}{\binom{n}{K}},$$

where $n = 2517$ is the number of genes encoded in WH8102; K is the number of genes predicted to be in the network by a step/method in the protocol; V is the number of predicted genes that are verified by microarray data according our criterion (see below); and A is the number of genes affected by the experimental manipulation.

In addition, we have validated the predicted NtcA promoters/binding sites based on the assumption that genes that were differentially expressed under our experimental conditions are more likely to bear NctA promoters/binding sites than those whose expression levels were not significantly affected. Let $p(S(g) > s)$ be the probability that gene g bear a putative promoter/binding site with a score $S(g) > s$ [for the formulation of $S(g)$, see (14)]. We then used the following log odds ratio (LOR) function to measure the confidence of the prediction:

$$\mathrm{LOR}(s) = \ln\frac{p(S(g_a) > s)}{p(S(g_o) > s)},$$

where $S(g_a)$ is the score of the putative promoter/binding site found for gene g_a that was affected (up- or down-regulated) under our experimental conditions, and $S(g_o)$ is that of the putative promoter/binding site found for gene g_o that was not significantly affected under our experimental conditions.

Microarray fabrication

Microarray data was obtained using a whole-genome microarray constructed and utilized as described in (I. T. Paulsen, manuscript in preparation). Briefly, we constructed a complete genome microarray for WH8102, consisting of a mixed population of PCR amplicons (2142 genes) and 70mer oligonucleotides (389 genes). Unique PCR amplicons representing each gene are ~800 bp, or smaller if the gene size is smaller. Unique 70mer oligonucleotides were utilized for genes under 300 bp and for the two genes that we were unable to amplify by PCR. Six complete replicates of the 2526 member (2517 of these genes are annotated in NCBI) gene set are printed on aminosilane coated Corning ultraGAP glass slides and irreversibly bound by ultraviolet (UV) crosslinking. Each array slide also includes a variety of negative controls (50% DMSO / 50% deionized water) and positive controls (including a total mix of WH8102 PCR amplicons, spiked *Arabidopsis* PCR amplicons and 70mer oligonucleotides).

Cell culture

WH8102 cells were grown in a synthetic ocean water-based (SOW) medium. The SOW component was made according to Morel (38). This SOW was then used at 75% with 25% MilliQ ultrapure water (Millipore). Nutrient levels are similar to SN (39), but several changes have been made as detailed here. Phosphate is added to 87.6 μM. Vitamins are added from a f/2 vitamin stock (40) at final concentrations: vitamin B12 (0.1 nM), biotin (2 nM) and thiamine (300 nM). Sodium - carbonate is added at 96.8 μM. EDTA ($Na_2EDTA2H_2O$) is added at 13.45 μM. Trace metals are added from a 1000X trace metal stock to final concentrations: $ZnSO_4$ (772 nM), $MnCl_2$ (7.07 μM), $CoCl_2$ (85.9 nM), Na_2MoO4 (1.16 μM), Citric acid (30 μM), $FeCl_3$ (740 nM), Na_2SeO_3 (50 nM) and $NiCl_2$ (50 nM). The medium was autoclaved before use. For this study, cells were grown with either 1 mM ammonium chloride or 9 mM sodium nitrate as the nitrogen source. A total of 1 l cultures in glass flasks were stirred gently during growth at 25°C and ~50 μmol photons m^{-2} s^{-1}. Contamination was checked by microscope examination and by plating on agar-solidified SOW media as above or similar media with 5 g glucose l^{-1} and 0.5 g tryptone l^{-1} added.

Microarray hybridization and data analysis

Total RNA from exponential phase cells was extracted using a Trizol-based method and purified using a Qiagen RNeasy kit following manufacturer's instructions. An indirect labeling method is used to label cDNA, where cDNA is synthesized in the presence of a nucleoside triphosphate analog containing a reactive aminoallyl group to which the fluorescent dye molecule is coupled. Control and treated samples are labeled with Cy3 and Cy5, respectively, pooled and hybridized to the same array. In addition, reverse labeling was routinely performed to reduce dye-specific biases in signal intensity. Slides were hybridized with labeled cDNA and then scanned and analyzed using TIGR's SPOTFINDER, MADAM and MIDAS software (41). RNAs from five nitrate grown cultures and four ammonium grown cultures were used. Thirteen different hybridizations were carried out, of which six used different RNA pools and seven were replicates of these six experiments, either dye swapping experiments or direct replicates. Statistical analyses were carried out using the Significance Analysis of Microarrays (SAM) software package (42), and the thirteen hybridizations were treated as independent experiments. SAM orders the genes by using a modified t statistic called relative difference based on the ratio of change in gene expression to standard deviation in the data for that gene. It declares a gene to be up- or down-regulated if the difference (D-value) between the observed relative difference and expected relative difference is above (D-value > 0) or below (D-value < 0) the global cutoff point, respectively (42). This procedure allows estimation of the median of false discovery rates. In this study we selected the cutoff points to control this median to be at most 5%. We call that a gene was down- or up-regulated by ammonium if it has a negative or positive D-value beyond the selected cutoffs, respectively.

RESULTS AND DISCUSSION

Template networks

Nitrogen assimilation in cyanobacteria is mainly controlled by the global regulator NtcA (34). Some components of this network have been previously studied in detail in the following cyanobacterial species/strains: *Nostoc* sp. PCC 7120 (PCC7120), *Synechocystis* sp. PCC 6803 (PCC6803), *Synechococcus* sp. PCC 7942 (PCC7942) and *Synechococcus* sp. WH 8103 (WH8103). The known parts of the nitrogen assimilation networks in these species were constructed as parts of the template networks.

(i) *Nitrogen assimilation network in PCC7120*: we have collected 58 genes known to be involved in the nitrogen assimilation process in PCC7120 through literature searches (Supplementary Table S1) (34). The ABC transporter NrtABCD is responsible for taking up nitrate or nitrite, which is subsequently reduced to ammonium by the nitrate reductase NarB and/or the nitrite reductase NirA. Leaked ammonium or that in the environment is taken up by the ammonium permease Amt1. This organism can also utilize urea through the urea transporter UrtABCDE and the urease UreABCDEFG. In absence of combined nitrogen sources, a portion of PCC7120 cells can develop into a specialized cellular form called a heterocyst for nitrogen fixation (12,34). Nitrogen fixation (*nif*) genes are only expressed in heterocysts, and some are known to be under NtcA control, such as *nifHDK* (12). Some genes involved in the heterocyst differentiation such as *hetC*, *hetR*, *devBCA* and *xis* are also directly or indirectly regulated by NtcA (34). Ammonium is subsequently incorporated into the carbon skeleton through the glutamine synthetase (GlnA)-glutamate synthase (GS/GOGAT) cycle. The activities of some of the nitrogen assimilation genes/proteins are further regulated by the signal transduction protein P_{II} (product of *glnB*) (15) and the regulator NtcB (43). Most of these genes are up-regulated by NtcA (12,34). However, *rbcL* that encodes the large subunit of ribulose biphosphate carboxylase/oxygenase (44) and *gor* that encodes glutathione reductase (45) are repressed by NtcA.

(ii) *Nitrogen assimilation network in PCC6803*: we have collected 15 genes that are known to be involved in the nitrogen assimilation process in PCC6803 through literature searches (Supplementary Table S2) (34). As shown in Supplementary Table S2, the nitrogen assimilation network in PCC6803 is generally similar to that in PCC7120, but this organism is not capable of nitrogen fixation and differentiation into heterocysts as it lacks the required genes. In addition, the glutamine synthetase (type I), GlnA, is subject to inactivation by two small proteins GifA and GifB, whose genes are both repressed by NtcA (46). Moreover, an additional glutamine synthetase of type III, GlnN, is encoded in this genome. Most of the genes of these proteins are known to be regulated by NtcA (34).

(iii) *Nitrogen assimilation network in PCC7942*: Although the complete genome sequence of this organism is not yet available, we have collected 17 genes that are known to be involved in the nitrogen assimilation network (34). As shown in Supplementary Table S3, the network in PCC7942 is similar to that in PCC6803, but the former is able to utilize cyanate as the sole nitrogen source by the ABC type cyanate transporter CynABD and the cyanase CynS (47) while the latter is not.

(iv) *Nitrogen assimilation network in WH8103*: This organism is phylogenetically close to WH8102 and inhabits a similar ecological niche (48). Although its genome sequence is not available, a recent study has demonstrated that WH8103 encodes the nitrate reductase NarB, the nitrite reductase NirA, the large subunit of ribulose bisphosphate carboxylase/oxygenase RbcL, the glutamine synthetase GlnA and the nitrate/nitrite transporter NrtP that belongs to the major facilitator family (Supplementary Table S4) (48). It has been suggested that the genes of these proteins are under NtcA regulation in this species (48).

The initial model of the nitrogen assimilation network in WH8102

The columns 4 of Supplementary Tables S1–S4 show the mapping results of the components of the nitrogen assimilation networks in PCC7120, PCC6803, PCC7942 and WH8103 onto WH8102, respectively, by P-MAP (21) or BLASTP. Table 1 shows the result of the union operation of these mappings with conflicts being resolved through a detailed phylogenetic tree analysis. For example, though two different sets of template genes *nrtD*, *nrtB* and *nrtA* of both PCC7120 and PCC6803, and *cynD, cynB* and *cynA* of PCC7942 are mapped to *synw2485, synw2486* and *synw2487*, respectively, we have accepted the mapping from PCC7942 since our phylogenetic tree analyses suggest that they have the closest evolutionary distances (data not shown). The proteins in Table 1 constitute the components of the initial model of the nitrogen assimilation network in WH8102. This model is in good agreement with the findings reported by Palenik *et al.* (6) through a computational analysis that the WH8102 genome encodes the essential components required for utilizing nitrate/nitrite, cyanate, urea and ammonium as nitrogen sources. Moreover, unlike some other cyanobacteria such as PCC7120, WH8102 lacks the nitrogen fixation machinery. Furthermore, it lacks canonical ABC type nitrate/nitrite transporters as in PCC6803, PCC7120 and PCC7942. Instead, it encodes the NrtP transporter for nitrate/ntrite uptake, like some other marine cyanobacteria such as WH8103 (48) and *Synechococcus* sp. PCC 7002 (49). Fourteen out of 27 genes in the model were differentially expressed under our experimental conditions (Table 1). The possible reasons that the expression of the other genes was not affected will be discussed later. Since WH8102 lacks all the *nif* genes and other genes required for heterocyst development, and the predicted *dev* genes (*synw1085-1087*) (Supplementary Table S1) were not differentially expressed under our experimental conditions, it is currently unclear what the roles of the *dev* genes are in the nitrogen assimilation network in WH8102. Hence, these three genes are excluded from our initial model (Table 1). However we retained the *petH* gene in the initial model, since it was down-regulated by ammonium.

Expansion of the initial network model

In this step, we expand the initial model by adding additional genes based on our predictions of operons, protein–protein interactions and regulatory binding sites (regulon), and phylogenetic profile analysis.

Expansion based on operon predictions. Using the predicted operon information (19,20) (http://csbl.bmb.uga.edu/~zhx/pathways/nitrogen/), two new genes are recruited into the network model. Specifically, *synw0273* (hypothetical protein) and *synw0274* (hypothetical protein) are recruited by

Table 1. The components in the initial network model

Name	Protein ID	Synonym	Operon	Templates	SAM *D*-value[a]	Rank of NtcA binding site[b]
amt1	33864789	*synw0253*		PCC6803		9
glnA	33865607	*synw1073*		PCC7120, PCC6803	−3.820	3
glnB	33864998	*synw0462*		PCC7120, PCC6803	6.422	26
gor	33866067	*synw1533*		PCC7120		299
icd	33864702	*synw0166*		PCC7120, PCC6803		618
narB	33866994	*synw2464*		PCC7120, PCC6803, WH8103		616
nirA	33867007	*synw2477*		PCC7120, PCC6803, WH8103		11
cynA	33867017	*synw2487*	*synw2485–synw2487*	PCC7942	−5.132	4
cynB	33867016	*synw2486*	*synw2485–synw2487*	PCC7942		4
cynD	33867015	*synw2485*	*synw2485–synw2487*	PCC7942		4
cynS	33867020	*synw2490*		PCC7942	−3.420	735
nrtP	33866993	*synw2463*		WH8103	−5.184	960
ntcA	33864811	*synw0275*	*synw0273–synw0275*	PCC7120,PCC6803	−5.098	1
petH	33865285	*synw0751*		PCC7120	−6.842	162
rbcL	33866250	*synw1718*		PCC7120		1101
ureA	33866979	*synw2449*	*synw2446–synw2449*	PCC7120	−3.602	98
ureB	33866978	*synw2448*	*synw2446–synw2449*	PCC7120	−6.018	98
ureC	33866977	*synw2447*	*synw2446–synw2449*	PCC7120	−7.659	98
ureD	33866976	*synw2446*	*synw2446–synw2449*	PCC7120		98
ureE	33866975	*synw2445*	*synw2443–synw2445*	PCC7120		78
ureF	33866974	*synw2444*	*synw2443–synw2445*	PCC7120		78
ureG	33866973	*synw2443*	*synw2443–synw2445*	PCC7120		78
urtA	33866972	*synw2442*	*synw2438–synw2442*	PCC7120		2
urtB	33866971	*synw2441*	*synw2438–synw2442*	PCC7120	−4.551	2
urtC	33866970	*synw2440*	*synw2438–synw2442*	PCC7120	−2.995	2
urtD	33866969	*synw2439*	*synw2438–synw2442*	PCC7120	−4.429	2
urtE	33866968	*synw2438*	*synw2438–synw2442*	PCC7120	−4.423	2

[a] SAM D values, a positive or negative number indicates that the gene was up- or down-regulated by ammonium, respectively.

[b] The top 54 predictions are considered as putative NtcA promoters. The NtcA promoter for *nrtP* is probably missed by our program, see text.

synw0275 (*ntcA*) because they are predicted to form an operon (Table 1). In agreement with this, both *ntcA* and *synw0273* were down-regulated by ammonium. However, the expression of *synw0274* was not affected. One possible reason might be that our probe for *synw0274* was somehow not able to detect the change in its expression (for the other possible reasons, see below). Previous studies have suggested that *ntcA* is likely to be transcribed monocistronically in both PCC7942 (50) and PCC7120 (51). In agreement with these findings, the genomic context of *ntcA* in both PCC7942 (https://maple.lsd.ornl.gov/microbial/syn_PCC7942/) and PCC7120 strongly suggests that *nctA* in these two species each forms a single-gene transcriptional unit (unpublished data). Thus, the discrepancy between our data and the previous findings might be due to the difference among the structures of the *ntcA* transcription units in WH8102, PCC7942 and PCC7120. Hence, our prediction that *ntcA* forms an operon with *synw0274* and *synw0273* in WH8102 might be a unique feature of WH8102. We surmise that SYNW0273 and SYNW0274 might be candidate modulators of NctA.

Expansion based on predicted protein–protein interactions. We first tested our two methods for protein–protein interaction prediction by applying them to *E.coli* K12 for the purpose of validation. Through orthology mapping between the *E.coli* K12 proteins and the protein interaction maps of *H.pylori* and *S.cerevisiae*, we predicted 831 interactions in *E.coli* K12 with scores (see Materials and Methods) better than 4.0, among which 30 pairs are known interactions ($P < 10^{-30}$). Through protein fusion analysis, we predicted 1807 pair-wise interactions in *E.coli* K12 using a cutoff 4.0, of which 17 are known interactions ($P < 10^{-8}$). Moreover, the distribution of the confidence scores of the non-verified predictions is very similar to that of the verified predictions for both methods as shown in Figure 1A and B, respectively. These results strongly suggest that the predictions by both orthology mapping and protein fusion analysis are highly statistically significant and hence are likely to be reliable. By combining the results of the two prediction methods, we predicted 2645 interactions in *E.coli* K12 with 47 interactions having experimental verification ($P < 10^{-32}$). The distribution of the scores of the combined predictions is shown in Figure 1C.

Having demonstrated the reliability of these two prediction methods on *E.coli* K12, we applied them to WH8102. Through orthology mapping, we predicted 165 and 128 pair-wise interactions in WH8102 based on the known interactions in *H.pylori* and *S.cerevisiae*, respectively (cutoff = 4.0). Five interactions are common between the two sets. In addition, we predicted 680 interactions based on gene fusion analysis at a cutoff 4.0, among which five pairs are common with the predictions by the orthology mapping (288 interactions). The distribution of the scores of these predicted interactions is similar to that of our predictions in *E.coli* K12 by the same method (Figure 1A and B), suggesting that similar levels of prediction accuracy are likely achieved for WH8102 to those for *E.coli*. It is not surprising that the fractions of overlapping interactions predicted by different methods are rather small, if one considers that the same is true for interactions determined by different groups using even the same experimental procedure (52). One explanation could be that each method or group only predicted or determined a small fraction of the vast interactome of an organism, and that many of the true interactions are probably transient and their appearance and disappearance are highly sensitive to the cellular conditions.

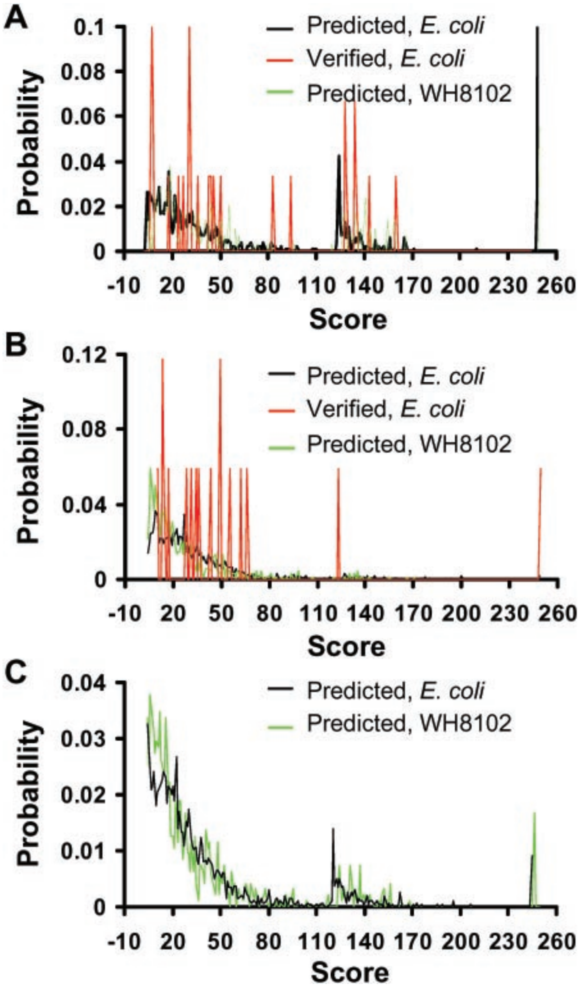

Figure 1. (**A**) Distributions of the confidence scores of the protein–protein interactions predicted by orthology mapping. (**B**) Distributions of the confidence scores of the protein–protein interactions predicted by protein fusion analysis. For both the methods, the distribution of the scores of the non-verified predictions in *E.coli* K12 (black lines) and that of the predictions in WH8102 (green lines) are very similar to that of the verified predictions in *E.coli* K12 (red lines). (**C**) Distributions of the confidence scores of the combined protein–protein interactions predicted by the two methods. The distributions of the scores of the predicted protein–protein interactions in *E.coli* K12 and WH8102 are very similar to each other, suggesting that similar prediction accuracy has been achieved for both the species.

By combing the predictions by the two methods, we obtained 950 pair-wise interactions with 713 proteins involved in WH8102. The distribution of their scores resembles that of the combined predictions for *E.coli* K12 (Figure 1C), suggesting again that the same level of statistical significance has probably been achieved for WH8102 as for *E.coli* K12 ($P < 10^{-32}$). The details of these predictions are available at http://csbl.bmb.uga.edu/~zhx/pathways/nitrogen/. A graphical representation of the predicted protein–protein interaction map is shown in Figure 2A. The distribution of the number of interactions per protein follows a power-law distribution [$C(n) = n^{-\gamma}$, where C is the number of proteins with

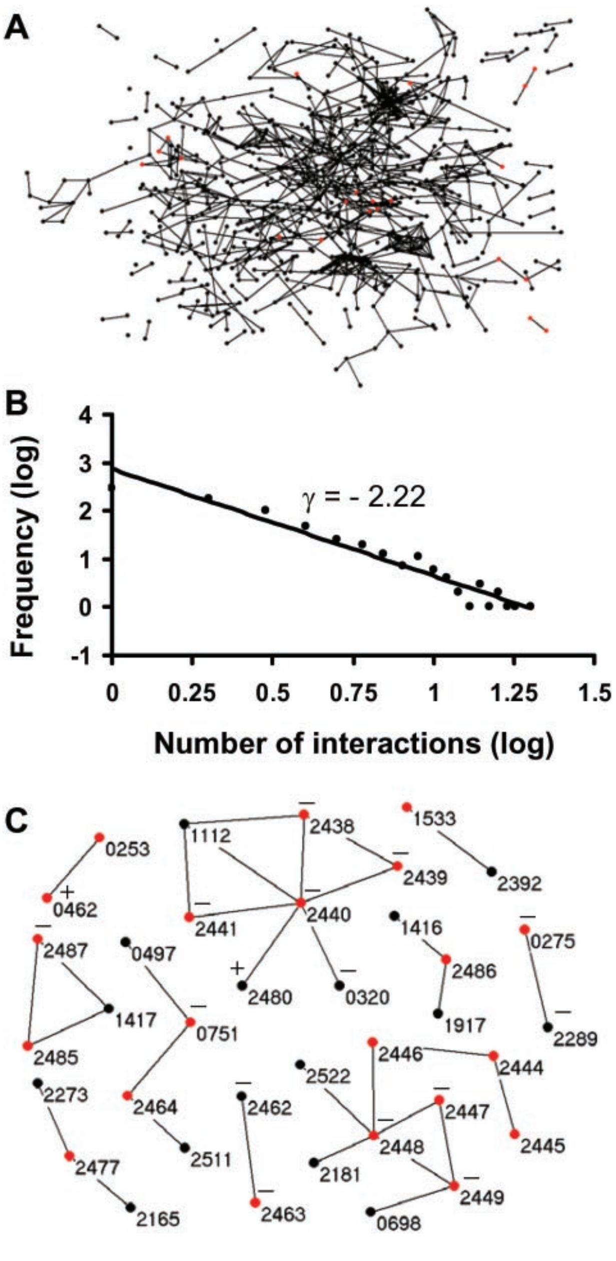

Figure 2. (**A**) Predicted genome-scale protein–protein interaction map in WH8102. Each vertex represents a protein, and an edge represents a predicted interaction. Vertices in red are proteins in the initial model of the nitrogen assimilation network. (**B**) The distribution of the degree of the vertices of the predicted protein–protein interactions in WH8102. It can be fitted to a power-law function, $C(n) = n^{-\gamma}$, where C is the number of proteins with n interacting partners and γ is a constant. (**C**) Proteins recruited into the network model through predicted protein–protein interactions. The vertices in red are the proteins in the initial model and those in black are proteins recruited. Proteins marked by '+' or '−' were up- or down-regulated by ammonium, respectively.

n interacting partners and γ is a constant] with $\gamma = 2.22$ (Figure 2B), which is typical for biological networks (53).

Using this interaction map, we have recruited additional proteins into our network model. Figure 2C shows in detail the predicted interacting partners of the components in the initial model. It is interesting to note that some known functionally related proteins in the initial network model are predicted to form complexes among themselves and with others that in some cases were differentially expressed under our experimental conditions (Figure 2C), suggesting that our prediction of protein–protein interactions makes biological sense. To name few such cases, the urea transporter subunits SYNW2438-2441 (UrtEDCB) form a complex with SYNW1112 (putative ABC transporter), SYNW0320 (putative ABC transporter) and SYNW2480 (putative ABC transporter); and *synw0320* and *synw2480* were down- and up-regulated by ammonium, respectively (Figure 2C). The urease subunits SYNW2444-2449 (UreFEDCBA) form a complex with SYNW2181 (possible exodeoxyribonuclease), SYNW2522 (excinuclease ABC subunit A) and SYNW0689 (hypothetical protein) (Figure 2C). The cyanate transporter subunits SYNW2485 (CynD) and SYNW2487 (CynA) form a complex with SYNW1417 (putative ABC transporter) (Figure 2C). In addition, the ammonium transporter SYNW0253 (Amt1) is predicted to interact with the signal transduction protein SYNW0462 (P_{II}, product of *glnB*) (Figure 2C), suggesting that Amt1 is likely to be regulated by P_{II} through a physical interaction as has been shown in *E.coli* (54,55). We have also predicted that the nitrate transporter SYNW2463 (NrtP) interacts with SYNW2462, both of which were down-regulated by ammonium (Figure 2C). Interestingly, the orthologs of SYNW2462 and SYNW2463 fuse into a single peptide in PCC7002 (49) and WH8103 (48). Moreover, the nitrogen regulator NtcA (SYNW0275) is predicted to interact with another response regulator SYNW2289, and both were down-regulated by ammonium (Figure 2C), suggesting a possible mechanism for the cross-talk between the nitrogen assimilation pathway and the pathways under the control of SYNW2289. Overall, 16 new proteins are recruited into the network by this method (Figure 2C). However, for most of these predictions, the functional relationships between the recruited proteins and the recruiting ones are largely unknown, which warrant further experimental investigations.

Expansion based on NtcA promoter predictions. In the current study, we have made a new prediction of the NtcA regulon in WH8102 through using two additional cyanobacterial genomes (*Synechococcus* sp. CC9605 and CC9902) in our analysis, which has improved the prediction accuracy of our previous study (14). The whole list of the predicted NctA promoters in WH8102 is available at http://csbl.bmb.uga.edu/~zhx/nitrogen/ntca/. In the current study, we predicted 54 transcriptional units (containing 102 genes) bearing NtcA promoters with $P < 0.05$ (at score cutoff of 11.166) (Table 2), of which 13 genes are already in the initial model, i.e. *synw0275 (ntcA), synw2442-2438 (urtABCDE), synw1073 (glnA), synw2487-2485 (cynABD), synw0253 (amt1), synw2477 (nirA)* and *synw0462 (glnB)*. We thus recruited 89 new genes into the nitrogen assimilation network (Table 2) by this method. Interestingly, the orthologues of the following recruited genes are known to play roles in the nitrogen assimilation networks in other cyanobacteria or bacteria: the type II alternative RNA polymerase σ factor SYNW2496 (RpoD) (56), the sodium/glutamate symporter SYNW0882 (GltS) (57), the porin SYNW2224 (Som) (16) and the molybdenum cofactor biosynthesis proteins SYNW2468 (MoaC) and SYNW2469 (MoeA) (Table 2). Interestingly, some of the recruited genes are involved in photosynthesis, such as

Table 2. The components of the network recruited by NtcA promoter predictions at $P < 0.05$

Rank	Transcription unit[a]	Names	NtcA site	Downstream of NtcA binding site and −10 like box[b]	NtcA site position	Score
1	***synw0275*** *synw0274* ***synw0273***	*ntcA* - -	GTAataactacTAC	ACGGCTCGTCAGGGCACATTCG**TAATTT**CCC	−49	14.518
2	*synw2442* ***synw2441*** ***synw2440*** ***synw2439*** ***synw2438***	*urtA1 urtB urtC urtD urtE*	GTTccggttgaTAC	CAAAGCGGTGGGGGGCCCTTTTT**TACCTT**CC	−52	14.188
3	***synw1073***	*glnA*	GTGcgcgttgaTAC	AAAACAGGGCATAACGGCTCCT**TACGGT**CGT	−60	13.978
4	***synw2487*** *synw2486 synw2485*	*cynD cynB cynA*	GTAtcacctgaTAC	AACATCCGCGTTCGCTTTCCAAC**TATAAA**TA	−51	13.831
5	*synw0165*	-	GTAgtttaggaAAC	ATATGGGTTAAAGATTTTTCGT**TATCAA**GAG	−39	13.152
6	*synw1105*	-	GTGttagttaaTAC	ACAAGCATGTACTAGACTGCGC**TAGTTT**AAT	−64	13.112
7	*synw0153 synw0154*	- -	GTAgtcgccgcTAC	ATCTGGTGGGGTGGGCAGACCG**TCCTCC**ACC	−62	13.111
8	***synw0347***	-	GTAgctaatttTAC	TCTAGCCTTGTTTTCTATATAGC**TAGCAC**TA	−173	13.097
9	*synw0253*	*amt1*	GTTcagtcggaTAC	ACCATCCGGCGTGACCAGCAGCTC**TGCACT**C	−37	13.015
10	***synw1434*** *synw1435*	- -	GTAacaacaccTAC	AGCCTGAACCAGTCCACTCGG**TAACAC**TATG	−143	12.792
11	*synw2477*	*nirA*	GTAattccatcAAC	AGAACAACTTTTGAGTACGAAC**TAGAAA**AGG	−349	12.711
12	***synw1412*** *synw1413 synw1414* ***synw1415*** *synw1416 synw1417 synw1418*	- *hypA2 hypB* - - - -	GTAgctgatcaCAC	CGCGCGTGCCACCGGTGCCGCA**GACAGT**GGA	−69	12.678
13	*synw2476 synw2475*	- *cobA*	GTTgatggaatTAC	GATTCCGCTCGTATTGCCGTCTG**TAACTC**TT	−199	12.586
14	*synw1508 synw1507*	- -	GTAataaagacTGC	GGAATTAATATTTCGGCAATACTTA**TACCTT**	−111	12.538
15	*synw0152*	*rnc*	GTAgcggcgacTAC	CGATATCGGCGCTCCTGACGGGGC**TGGCGG**G	−527	12.477
16	***synw1426*** *synw1425 synw1424 synw1423 synw1422*	- - - - -	GTCgtatttcgTAC	ATTTTTTGTGGGCCGACCGGAGCCAGTCTTT	−52	12.416
17	*synw2171*	-	GTCatggatacTAC	CCTTGCCCACCTCTGTACACTT**TCGGGT**AGC	−56	12.168
18	***synw1427***	-	GTGgccttggcTAC	AACACAAGTGTTGATTTCAAAC**AAGGAT**TAG	−56	12.160
19	***synw1222*** *synw1221*	*metG* -	GTCccaggtcaTAC	GGGCAACAAAACAGAAAAG**TGCGGT**CGTTAG	−72	12.098
20	***synw2492***	-	GTAtcaaccacTAC	TTTCATGGGGTCCGGCAC**TCCTGA**ACACTTC	−53	11.997
21	***synw2524*** *synw2525 synw2526*	- - *thrC*	GTGgtgatgccGAC	CGTAGCGACGGCTCGCGTCGTT**TAGAAG**GGG	−32	11.987
22	*synw0768 synw0769*	- -	GTCgcagcaggGAC	CGCCAGATTGGCATCCATGGCCCC**TAACAA**C	−251	11.976
23	*synw0433 synw0434 synw0435 synw0436 synw0437*	- - *hisH* - -	GTTatagccaaTAT	ATAATTTTAGCTTATTCACGGA**TAAATT**TTG	−141	11.947
24	***synw2151***	*psbA3*	GTAaacgggcgAAC	ACACCTCTCAACCTTCCGT**TACATT**GGCGTG	−93	11.936
25	*synw0346*	-	GTAaaattagcTAC	AGAACAAAAGCAATTGGCT**TAATTG**ATACCA	−627	11.920
26	*synw0462*	*glnB*	GTTacaggggcTAC	CCACACCGCCACCATTCACG**TCATGC**TTAAT	−51	11.915
27	***synw2097***	-	GTGctcagcgtTAC	CAAGGGGGGCTTGGTGGAATCGA**TAGCAT**GC	−51	11.883
28	*synw0015*	-	GTGgtattccgTAC	GGGGAATGATCAATTTGAAATC**TACTTT**CGT	−423	11.865
29	*synw0299*	-	GTCcgcttcgcCAC	CGGCATGACACACCATGGGG**TAAAAG**GGAAG	−51	11.816
30	*synw0267*	-	GTCaggcgtaaTAC	CGCCAAAGACTTTCATCAGAGGGC**TGCGAT**C	−222	11.787
31	***synw2456*** *synw2457* ***synw2458***	- - -	GTCgctggcgcTAC	CCAAATCACAGGCTGTCAGCGCAC**TGATGC**C	−89	11.753
32	*synw1544*	*futC*	GTAcagccagcCAC	CGGCCAGCACCGCCTGGTTG**TAAGGC**CCCAC	−71	11.743
33	*synw1264*	-	GTAaagaccgcTAC	CGCCAATGACACCAACACGGGCGT**TGCTGA**G	−80	11.615
34	*synw2185*	-	GTAcgcgtaggAAC	ATTGAGAGGATTTCTCCCTAG**TACGAG**AGGA	−662	11.614
35	*synw0463 synw0464*	- -	GTAgcccctgtAAC	CGAATGAGTGCCTTCGCCTTC**TAGCGT**TCTG	−70	11.564
36	*synw2496 synw2497*	*rpoD* -	GTGatgactgaTAC	GGAATCGTTGCAATCAAGC**TGAAGC**CGTTCA	−159	11.563
37	*synw2482 synw2483*	- -	GTCaaatacgaTAC	GAAAAACCAAAGATATAT**TTGAAT**CACTGAA	−91	11.533
38	*synw1703 synw1704 synw1705 synw1706*	- *cobB* - *B40* -	GTCtgagacggAAC	ACGGCTCAGCCCGAGGCTGACTGA**TCCATT**T	−628	11.526
39	*synw1273*	-	GTAtttgtataTAC	ATTATCTTTGGCATCGGTTGG**TGCCTT**TGCC	−398	11.480
40	*synw1274 synw1275 synw1276* ***synw1277***	*petF2* - - -	GTAtatacaaaTAC	TCCCTCGACCCGAGACATCGAAT**TATGGA**CT	−391	11.432
41	*synw1662*	-	GTCtttttacaTAC	GTAATGTTATTTGGATGGGC**TCCGTT**TGAGG	−452	11.380
42	*synw0882*	*gltS*	GTGagcctcaaTAC	TTACGGTACTTTGATGGC**TAGGAT**CTCGAAA	−151	11.356
43	*synw0862*	-	GTCttgattgaTAC	GGATGGCTTGAGCTCGAGTCG**TCAGCT**TGAT	−103	11.318
44	*synw1820 synw1819* ***synw1818***	- - -	GTGactgaggcTAC	TGCTGTAGAGCCGTCGTCCAAC**TTCCGT**CAG	−704	11.288
45	*synw2417 synw2418*	- -	GTAaaaaagcgGAC	TTGCGACCGCTCAGAAACGAGAAG**TGAACT**G	−177	11.273
46	*synw0118*	-	GTCgagctcctTAC	GGCCGAAGGCGGCCTGAT**TGATGT**CGGCAAT	−113	11.257
47	*synw2255*	-	GTGtcaacgatGAC	GGTCCGACACCGACGAATCGCATT**TCCCTT**C	−37	11.236
48	*synw2466 synw2467 synw2468 synw2469*	- - *moaC moeA*	GTTcatctcgtTAC	ACATCCACGTCCGTGCGATTCGCCC**TCTCCC**	−209	11.232
49	***synw2175***	-	GTAaccgccgcCAC	CGGCCCAGCTGAAGAACAACAAA**TAACTC**AC	−130	11.211
50	*synw2113*	-	GTTcccggctgCAC	AAGGTCGGGAGCCAGCAGGCTGCT**TAACGT**C	−53	11.203
51	***synw2150***	-	GTTcgcccgttTAC	AAAGCAAATGTGAAGGATGGT**TGAGCG**CCGT	−77	11.201
52	*synw1074*	*apcF*	GTAtcaacgcgCAC	ATTTTTTAATGTGCCTTTG**TTGGTT**CAGGGA	−149	11.175
53	*synw0164 synw0163*	- -	GTTtcctaaacTAC	AAGCTAATTTTGTGATAAACCTT**TATGAA**AA	−286	11.172
54	*synw2224*	*som*	GTAgtaatggaCAC	ATAAAAGCGCGGCCTCCTG**TGCAGT**GTTTTC	−94	11.166

[a] Bold face or underlined, down- or up-regulated when grown on ammonium relative to nitrate, respectively.

[b] Bold face, putative −10 like box.

the photosystem II D1 protein SYNW2151 (PsbA3), the ferredoxin SYNW1274 (PetF2) and the phycobilisome core component allophycocyanin β-18 subunit SYNW1074 (ApcF), suggesting a possible coupling between nitrogen assimilation and photosynthesis. In addition, a number of the recruited genes encode hypothetical proteins, which are likely to play roles in nitrogen assimilation related biological processes. Twenty-five and seven out of these 102 genes, predicted to be regulated by NtcA, were down- and up-regulated by ammonium, respectively (Table 2). Interestingly, the 20 transcription units containing the 25 genes down-regulated by ammonium generally bear a canonical NtcA promoter consisting of a pseudo-palindromic $GTAN_3TAC$ motif and a TAN_3T/A box located ~22 bases downstream of it (Table 2). In contrast, the five transcription units containing the seven genes up-regulated by ammonium only bear a GTA/GN_3TAC motif (Table 2), their predicted downstream −10 like boxes largely deviate from the consensus TAN_3T/A in all the cases (Table 2), suggesting that it might not be required for the negative regulation by NtcA. This result is consistent with previous observations (34).

Expansion based on phylogenetic profile analysis. We clustered proteins based on the similarity of their phylogenetic profiles. The idea is to derive co-evolved proteins through identifying proteins with similar phylogenetic profiles. Figure 3A is the 2D representation of the clustering results of the proteins in WH8102. As we have discussed in the Materials and Methods section, each valley in this plot corresponds to a cluster of proteins with similar phylogenetic profiles. We have identified a total of 197 clusters, among which 90 clusters have a statistical significance $P < 0.05$. The whole list of clusters is available at (http://csbl.bmb.uga.edu/~zhx/pathways/nitrogen/). A number of interesting features in these clusters are worth noting. The largest cluster with 443 proteins, labeled by the horizontal bar I, contains proteins unique to WH8102. The next seven clusters, labeled by the horizontal bar II, contain proteins unique to the 11 cyanobacterial genomes used in our analysis (http://csbl.bmb.uga.edu/~zhx/pathways/nitrogen/). The cluster labeled by the horizontal bar III contains 20 proteins shared by all the 231 genomes used in our analysis. These proteins are involved in the universal cellular functions, such as DNA replication, transcription and protein translation.

The mapping coordinates of the proteins of the initial network model are indicated by the open circles in Figure 3A, which are vertically separated for the clarity of presentation. If at least one protein in the initial network is in a cluster, we add all the other members of the cluster into the model. We have identified three such clusters shown in Figure 3B–D. Specifically, the cluster containing SYNW2443 (UreG), SYNW2449 (UreA), SYNW2448 (UreB) and SYNW2447 (UreC) already has all its members in the initial model ($P < 0.0175$), and the genes of the last three proteins were down-regulated by ammonium (Figure 3B). SYNW2464 (NarB) recruits four proteins into the network model, i.e. SYNW2469 (molybdenum cofactor biosynthesis protein), SYNW2460 (molybdenum cofactor biosynthesis protein), SYNW2468 (molybdenum cofactor biosynthesis protein C, *moaC*) and SYNW2474 (molybdenum cofactor biosynthesis protein B1, *moaB1*) (Figure 3C) ($P < 0.0448$). Again, the cluster containing SYNW2438

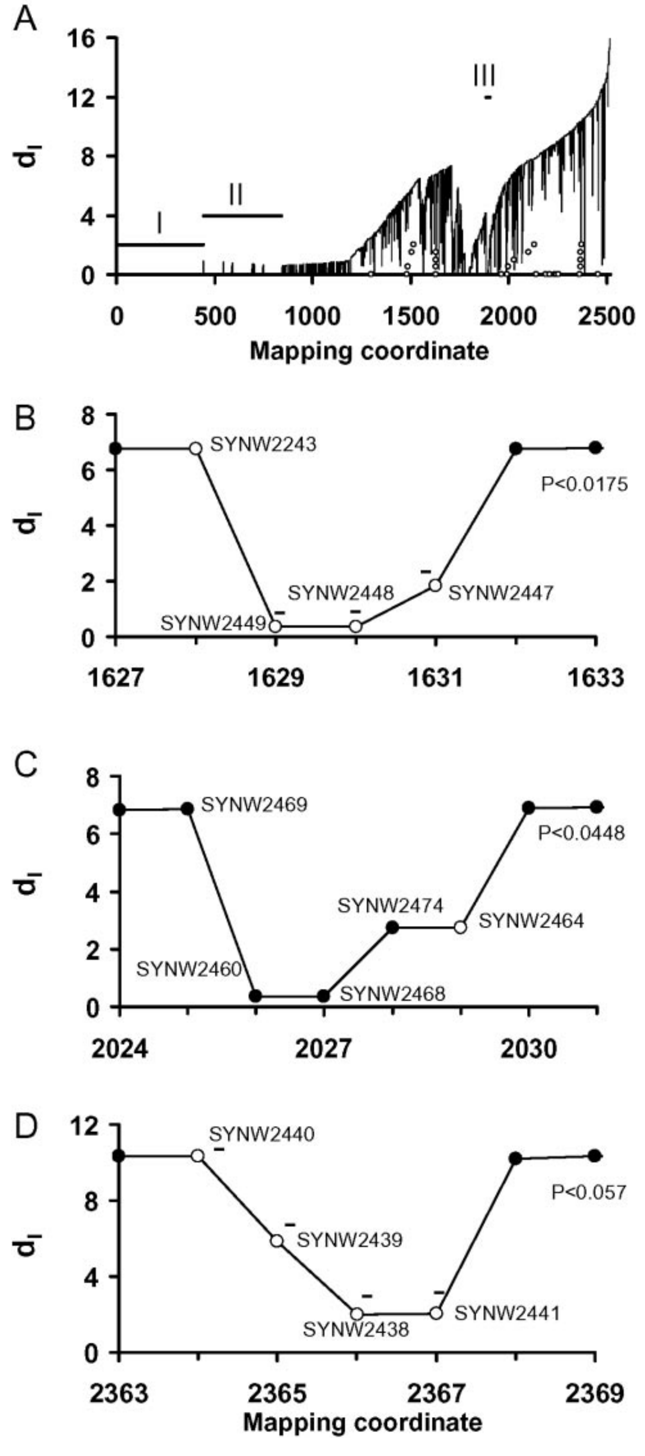

Figure 3. Phylogenetic profile analysis. (**A**) 2D representation of the clusters of the proteins of WH8102 based on the similarity of their phylogenetic profiles. The open circles near the horizontal axis indicate the mapping coordinates of the proteins in the initial model of the nitrogen assimilation network, and they are arbitrarily vertically separated to show the crowded ones. The cluster labeled by bar I contains proteins unique to WH8102. The next 7 clusters, labeled by bar II, contain proteins unique to the nine cyanobacterial genomes used in our analysis. The cluster III contains 20 proteins shared by all the 231 genomes in our analysis. B, C and D, close-up views to show the clusters containing at least one protein in the initial network model. The open circles represent proteins in the initial network. Proteins that were down-regulated by ammonium are labeled by '−'. The *P*-values indicate the statistical significance levels of the clusters.

(UrtE), SYNW 2439 (UrtD), SYNW 2440 (UrtC) and SYNW 2441 (UrtB) already has all its members in the initial model ($P < 0.0575$), and their genes were all down-regulated by ammonium (Figure 3D). The orthologues of a number of proteins in these clusters are known to either form a structural complex such as the urease by SYNW2443, SYNW2447-2449 and the urea transporter by SYNW2438-2441, or be functionally dependent such as the cluster that contains NarB and the enzymes responsible for the biosynthesis of the coenzyme factor of NarB, the molybdenum cofactor (Figure 3C). These results suggest that our clustering algorithm is able to cluster the proteins/genes that are functionally associated based on the similarity of their phylogenetic profiles, and that our phylogenetic clustering analysis makes good biological sense.

Validation and refinement of the predictions

To validate the predicted components in the model of the nitrogen assimilation network and refine it based on the validation results, we have compared our predictions with the whole-genome microarray gene expression data collected under ammonium-rich and nitrate-rich conditions (see Materials and Methods). We found that 247 genes were down-regulated, while 91 genes were up-regulated by ammonium based on the *D*-values from the SAM package (42) (see Materials and Methods). The genes that were down- or up-regulated by ammonium are summarized in Supplementary Tables S5 and S6, respectively. The raw data from the microarray experiments are available at http://csbl.bmb.uga.edu/~zhx/pathways/nitrogen/. It is highly likely that these 338 genes, either down- or up-regulated by ammonium, are involved in the nitrogen assimilation process or related global responses. It is worth noting that *synw0021* and *synw2225* are annotated as pseudo-genes in NCBI, but we found that they were both up-regulated by ammonium, indicating that they are likely to be functional genes (Supplementary Table S6).

We validated the genes recruited at each step/method of our prediction procedure by computing the probability of how likely the results of validation may happen by chance (see Materials and Methods). As shown in Table 3, high statistical significance levels have been achieved for the predicted components in our initial model ($P < 2.35 \times 10^{-7}$) as well as for the components recruited through prediction of operons ($P < 0.018$), prediction of NtcA promoters ($P < 1.55 \times 10^{-4}$), and prediction of protein–protein interactions ($P < 0.0525$). Although none of the four new genes, *synw2469, synw2460, synw2468* and *synw2474,* recruited through phylogenetic profile analysis was differentially expressed under our experimental conditions (Figure 3C), it seems to make biological sense to have them in the network model as they form a highly functionally related cluster with their recruiting gene *synw2464* (*narB*) (Figure 3C). The possible reasons that the genes in this cluster were not differentially expressed will be discussed later. In addition, if we include the cases where a gene in the initial model recruits another that is already in the initial model, then a much higher statistical significance levels are achieved for all the recruiting methods (Table 3). Overall, 42 out of 133 genes recruited into the network by our computational methods were differentially expressed under our experimental conditions with $P < 5.68 \times 10^{-9}$ (Table 3), suggesting that the network model constructed by our protocol is highly statistically significant, and thus is likely to be biologically meaningful. The whole list of the predicted genes of the network is given in Supplementary Table S7.

Table 3. Validation of the components of the nitrogen assimilation network recruited by the methods of the computational protocol

Methods	Number of[a] genes recruited	genes affected	*P*-value[a]
Initial model	27	14	2.35×10^{-7}
Protein-protein interactions	16 (32)	4 (14)	0.0525 (4.02×10^{-6})
Phylogenetic profile	4 (12)	0 (7)	–(2.96×10^{-5})
Regulon prediction	89 (102)	24 (32)	1.55×10^{-4} (3.89×10^{-7})
Operon	2 (17)	1 (9)	0.018 (1.36×10^{-5})
Combined	133	42	5.68×10^{-9}

[a] The number in a parenthesis is the result when the genes in the initial network model are also considered for the method, see text.

We have also validated our regulon prediction by computing the LOR, (see Materials and Methods) of the scores of the predicted NtcA promoters for the genes that were down-regulated over those that were not significantly affected under our experimental conditions. As shown in Figure 4A, the *LOR* increases smoothly beyond a certain score of the putative NtcA promoters, indicating that genes bearing high-scoring NtcA promoters were more likely to be down-regulated by ammonium than those that were not significantly affected. This result again suggests that our gene recruitment through NtcA regulon prediction is highly biologically meaningful. On the other hand, this result also suggests that the genes that were down-regulated by ammonium but do not bear high-scoring NtcA promoters, might not be directly regulated by NtcA. Using a score cutoff of 11.166 for the NtcA promoter predictions at $P < 0.05$ (14), we identified 222 such genes including six genes in the initial network model, i.e. *cynS (synw2490), nrtP (synw2463), ureA (synw2449), ureB (synw2448), ureC (synw2447)* and *petH* (*synw0751*) (Table 1). However, *nrtP (synw2463)* probably bears an NtcA promoter, but our program did not predict it due to the fact that this gene lacks orthologues in the other genomes (14).

We have also conducted a similar analysis on the genes that were up-regulated by ammonium and thus are likely to be negatively regulated by NtcA. In this case we only searched for the canonical pseudo-palindromic NtcA binding sites in the upstream regions of the predicted transcription units in WH8102, since it has been shown in PCC7120 and PCC6803 that genes under negative control of NtcA do bear an NtcA binding site, but not the -10 TAN$_3$T/A box located 22 bases downstream of it (34). The whole list of the predicted NtcA binding sites (without -10 like boxes) are available at http://csbl.bmb.uga.edu/~zhx/pathways/nitrogen. As shown in Figure 4B, the *LOR* goes above zero with the increase of the scores of predicted NtcA binding sites, suggesting that genes up-regulated by ammonium are more likely to bear canonical NtcA binding sites than those that were not significantly affected. Hence, at least some of the genes up-regulated by ammonium might be mediated by NtcA. However, as shown in Figure 4B, the *LOR* values are

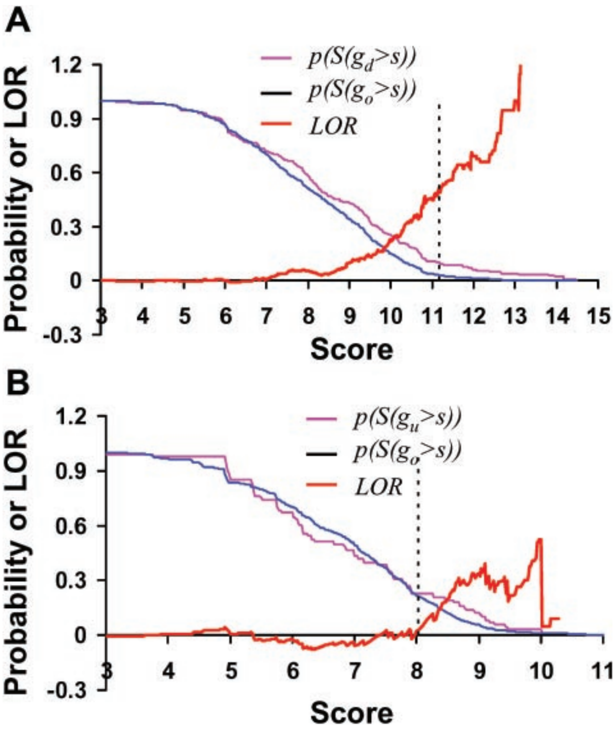

Figure 4. (**A**) The cumulative probability functions of the scores of putative NtcA promoters (an NtcA binding site plus a downstream TAN_3T/A box) found for genes down-regulated [$p(S(g_d) > s)$, pink] and for genes unaffected [$p(S(g_o) > s)$, blue] by ammonium, and their log odds ratio function (*LOR*, red, see Materials and Methods). The solid vertical line indicates the score cutoff 11.166 for the NtcA regulon prediction with $P < 0.05$. (**B**) The cumulative probability functions of the scores of putative canonical NtcA binding sites found for genes up-regulated [$p(S(g_u) > s)$, pink] and for genes unaffected [$p(S(g_o) > s)$, blue] by ammonium, and their log odds ratio function (*LOR*, red). The solid vertical line indicates the score cutoff 8.02 for predicting the canonical NtcA binding sites that are possibly involved in the repressive regulation by NtcA (see text).

not very high and oscillate as the score increases, suggesting that the prediction accuracy is not very high. To reduce the rate of false predictions, we consider only the predicted NtcA binding sites with the downstream genes up-regulated by ammonium as possible NtcA binding sites. Using a cutoff 8.02 where the *LOR* goes above zero (Figure 4B), we have predicted that 20 out of the 91 genes that were up-regulated by ammonium, bear canonical NtcA binding sites (Supplementary Table S6), which include *glnB* (*synw0462*), *som (synw2224)* and *rpoD (synw2496)*, the principle σ^{-70} factor (*synw1783*), and the photosystem I P 700 chlorophyll a apoprotein subunits Ib *psaB* (*synw2123*) and Ia *psaA* (*synw2124*).

Since we are not able to detect common elements in the upstream regions of the genes that were either down- or up-regulated by ammonium, yet lack putative NtcA promoters (or binding sites) (data not shown), we suspect that they are probably regulated by different regulators downstream of the NtcA regulon. Interestingly, we found that six (*synw0926, synw2401, synw1592, synw1875, synw0808* and *synw2289*) and two [*synw1462* and *synw0549* (putative circadian regulator)] putative specific transcription regulators were down- and up-regulated by ammonium, respectively (Supplementary Tables S5 and S6); hence we surmise that some of these genes might be mediated by these regulators. However, we could not provide stronger evidence for this hypothesis due to the lack of microarray data under multiple growth conditions.

It should be noted that though our predictions are highly statistically significant, we have verified only 31.6% (42/133) of the genes predicted to be parts of the nitrogen assimilation network, using our microarray data. Several possible reasons might account for this low number. First, our current experimental conditions (ammonium versus Na_2NO_3 as the sole nitrogen source) might represent a mild perturbation to the whole nitrogen assimilation network, thus only a portion of the network was affected. For instance, under a mild perturbation, NtcA might not be fully activated or repressed, and thus genes with a weaker promoter might not be affected. This might be one of the reasons that some predicted members of the NtcA regulon were not differentially expressed when the cells were grown on ammonium versus nitrate, including *amt1* (*synw0253*), *nirA* (*synw2477*), *gltS* (*synw0882*) *apeF*(*synw1074*) and *petF2* (*synw1274*), etc. (Table 2). These genes might be constitutively expressed at similar levels on ammonium and nitrate. We expect that more genes will be up- or down-regulated under stronger perturbations, such as nitrogen starvation or rapid ammonium addition to nitrate cultures. Second, some genes in the network may not necessarily change their expression levels under our experimental conditions due to their lack of relevant regulatory binding sites. For instance, no NtcA binding site is found for *narB* (nitrate reductase) in WH8102, thus, it is very likely to be true that the expression of *narB* was not affected in our experiments. Of course, the activities of possible regulators of *narB* were probably not affected under our experimental conditions either. Third, our microarray techniques might not be sensitive enough to detect subtle changes in expression of some genes. For instance, we predicted that *urtA1-urtB-urtC-urtD-urtE* (*sywn2442-2438*) form an operon with a high-scoring NtcA promoter in the upstream region of *urtA1*. Our microarray data shows that the expression of *urtB, urtC, urtD* and *urtE* was down-regulated by ammonium (Table 2), but that of *urtA1* was not. Since we did not find any NtcA binding site in front of *urtB*, we strongly believe that the expression of *urtB, urtC, urtD* and *urtE* is mediated by the NtcA promoter in front of *urtA1*. Thus, our failure to detect the change in *urtA1* expression is probably due to the technical limitations of our experiments and/or our data analysis programs. Another possibility is that the part of the polycistronic mRNA encoding *urtA1* might be more labile than the rest of the polycistronic mRNA encoding *urtB, urtC, urtD* and *urtE*. Finally, it is likely that we have made some false positive predictions when recruiting genes into the network model. However, it is currently difficult to estimate the false positive rate of our predictions due to the lack of an experimentally verified gene set for the network in WH8102. Nevertheless, based on our statistical estimation, we believe that the false positive rates are not very high, in particular for the NtcA regulon prediction (14).

It might be reasonable to include the genes that were down- or up-regulated by ammonium into our model of the nitrogen assimilation regulatory network, if they are not in our expanded model yet. Thus, we have further expanded our network model by recruiting 211 and 85 new genes that were down- or up-regulated by ammonium, respectively (Supplementary

Tables S5 and S6). For most of these genes, we do not know their precise roles in the nitrogen assimilation network, except for those bearing putative NtcA binding sites in their regulatory regions, which might be a part of the NtcA regulon negatively regulated by NtcA (Supplementary Table S6). Nevertheless, these genes might be responsible for local or more global responses invoked by the nitrogen regimes of our experiments.

Coordination between the nitrogen assimilation and photosynthesis processes

In agreement with our prediction that the photosynthetic gene *psbA3* (*synw2151*) is a member of the NtcA regulon, our microarray data indicates that *psbA3* was down-regulated by ammonium. In addition, an NtcA binding site is found for the putative operon *synw2124-synw2123*, which encodes the photosystem I P700 chlorophyll a apoprotein subunits Ia (PsaA,SYNW2124) and Ib (PsaB, SYNW2123) (Supplementary Table S6). Both *synw2124* and *synw2143* were up-regulated by ammonium, suggesting that this operon might be negatively regulated by NtcA through the predicted NtcA binding site (Supplementary Table S6).

Intriguingly, many photosynthetic genes lacking a canonical NtcA promoter were down-regulated by ammonium, such as the ferredoxin thioredoxin reductase catalytic β chain gene (*synw0318*), the cytochrome b6/f complex subunit (Rieske iron-sulfur protein) gene (*synw1841*), the photosystem II chlorophyll-binding protein CP43 gene (*synw0676*), the photosystem II D1 protein form II gene (*synw0983*), the 3Fe-4S ferredoxin gene (*synw0624*) and the high light inducible protein genes (*synw2403* and *synw0330*) (Supplementary Table S5). On the other hand, some other photosynthesis related genes were up-regulated by ammonium, but do not bear canonical NtcA binding sites in their regulatory regions, including the C-phycoerythrin class II chain genes (*synw2008* and *synw2010*), the C-phycoerythrin class I α chain gene (*synw2016*), and the light-independent protochlorophyllide reductase chlB genes (*synw1723*, *synw1724* and *synw2016*) (Supplementary Table S6). Thus, these photosynthetic genes that were either down- or up-regulated by ammonium, but lack NtcA promoters/binding sites are probably indirectly regulated by the downstream events of NtcA, presumably by some of the eight putative transcription regulators (*synw0926, synw2401, synw1592, synw1875, synw0808, synw2289*, *synw1462* and *synw0549*) that were down- or up-regulated by ammonium (Supplementary Tables S5 and S6). Taken together, these results unequivocally demonstrate that some photosynthetic genes are part of the nitrogen assimilation network.

As a matter of fact, it has been shown that nitrogen assimilation and photosynthesis are two highly coupled processes (58) though the molecular basis of the coordination between photosynthesis and nitrogen assimilation is largely unknown. Our demonstration that some photosynthetic genes are part of the nitrogen assimilation network in WH8102 might provide a strong evidence and possible mechanism for the coordination between the nitrogen assimilation and photosynthesis processes, and confirmed our previous analyses based on NtcA regulon predictions in nine sequenced cyanobacteria (14).

Although the biological significance is relatively well understood of why some nitrogen assimilation related genes are down-regulated while the others are up-regulated under different nitrogen regimes (34), it is not clear why some photosynthetic genes or those that are not directly related to nitrogen assimilation, are down-regulated, while the others are up-regulated under the same nitrogen regimes. It may be related to the global adaptation responses of a cell to the environmental changes as has been shown for the nitrogen starvation-induced chlorosis in PCC7942 (16–18), where the up- and down-regulations of photosynthetic genes are important for the long term survival of the cell under prolonged nitrogen starvation (16–18).

The working model of the nitrogen assimilation network

We now describe a working model for the nitrogen assimilation network in WH8102 based on the 133 genes predicted by our computational protocol (Supplementary Table S7) and the 338 genes revealed by the microarray data (Supplementary Tables S5 and S6) as shown in Figure 5, which consists of a total of 429 genes (42 genes are common in the two sets). The genes that are predicted to bear NtcA promoters (regardless they were down-regulated by ammonia or not, see Table 2) or canonical NtcA binding sites (and were up-regulated by ammonium, see Supplementary Table S6) are considered to be members of the NtcA regulon (shown in green), while the others lacking such binding sites are considered to be non-NtcA regulon members (shown in blue). In this model, the nitrogen regulator NtcA (SYNW0275) is activated by a high level of 2-oxoglurate (2-OG), indicative of high C/N balance, or relatively lower availability of nitrogen supply in the environment. Since the 2-OG level is also determined by the reducing power and energy provided by photosynthesis, 2-OG might serve as a signal from photosynthesis to the nitrogen assimilation process. The core of the nitrogen assimilation network consists of the members of the NtcA regulon, which includes genes responsible for the uptake of various sources of nitrogen and their subsequent reduction and incorporation into the carbon skeleton, such as Nrtp (SYNW2462-2463), NirA (SYN2477), CynCBA (SYNW2486-2487), GlnB or P_{II} (SYNW0462), GlnA (SYNW1079), Amt1 (SYNW0253), UrtEDCBA (SYNW2438-2442), the glutamate transporter GltS (SYNW0882), the porin Som (SYNW2224) and the genes under NtcA control but also playing roles in the other functional pathways such as the photosynthesis pathway, which might serve as regulatory points to coordinate various cellular activities and the nitrogen assimilation process. The next layer of the network consists of proteins such as NarB (SYNW2464), UreGFE (SYNW2443-2445), UreDABC (SYNW2446-2449), CynS (SYNW2490), GOGAT (glutamate synthase), RbcL (SYNW1718) and Icd (SYNW0166), etc. which are not members of the NtcA regulon but are functionally relevant to the network. Some of these proteins are probably regulated by the regulators that are under NtcA regulation. However, it remains largely unknown how such indirect regulations occur since no NtcA promoter or canonical NtcA binding site is found for any putative transcription regulator (Table 2) except for the σ-factor RpoD (SYNW2496) (Table 2) and the principle σ-factor SigA (SYNW1783) (Supplementary Table S6). Nevertheless,

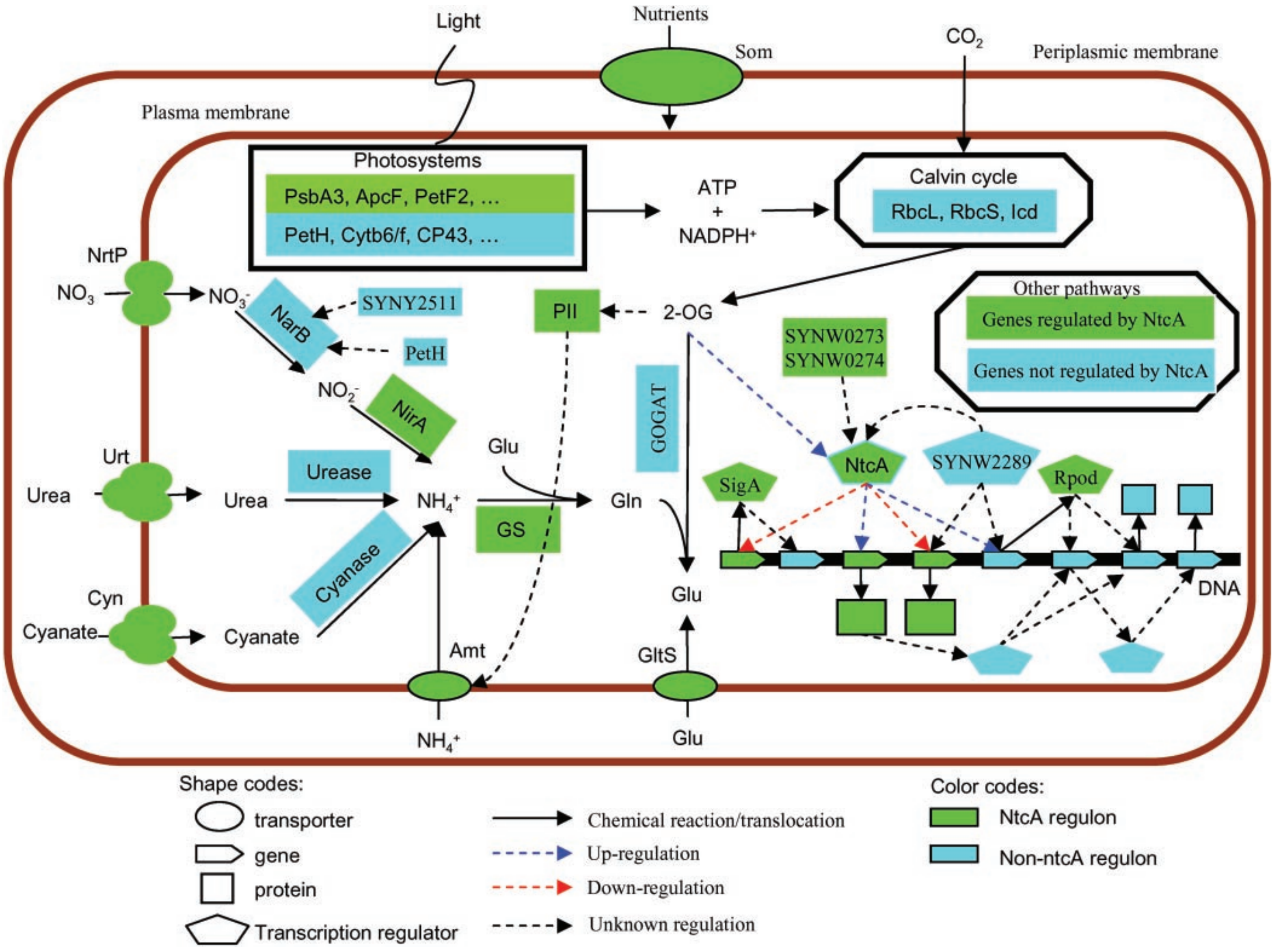

Figure 5. The working model of the nitrogen assimilation network in WH8102. Genes recruited either by our computational methods (Supplementary Table S7) or by microarray gene expression data (Supplementary Tables S5 and S6) are all considered to be members of the network. Genes that are predicted to bear NtcA promoters (Table 2) or canonic NtcA binding sites (also up-regulated by ammonium, Supplementary Table S6) are considered to be members of the NtcA regulon, while the others are considered to be non-NtcA regulon members of the network. Solid arrows represent substance translocations or chemical reactions, and dashed arrows represent regulatory relationships.

genes in the network may be up- or down-regulated directly by NtcA through NtcA binding sites/promoters, or indirectly by its downstream proteins under certain conditions (Figure 5).

Some other interesting predictions shown in the model are worth noting. The activity of NtcA is likely to be regulated by other proteins such as the response regulator SYNW2289 that is predicted to physically interact with NtcA, and the hypothetical proteins SYNW0273 and SYNW0274 that are predicted to be located in the same operon as NtcA. Amt1 is likely to be regulated by the P_{II} protein through a physical interaction as the reminiscent of a recent finding that the N-acetyl-L-glutamate kinase is regulated by the P_{II} protein through a physical interaction in *S.elongatus* (59), and this has been experimentally proven in *E.coli* (54,55). The function of the nitrate reductase NarB might be regulated by the ferredoxin-NADP reductase PetH (SYNW0751) and SYNW2511. The σ-factors RpoD (SYNW2496) and SigA (SYNW1783), which were down- and up-regulated by ammonium, respectively, are probably involved in the transcription of some genes that are down- or up-regulated by ammonium. Of the 429 genes recruited into the network, 204 are hypothetical genes; their possible functions in nitrogen assimilation related processes warrant further experimental investigations. While further experimental data are clearly needed to build a more detailed and accurate model, this working model should be very useful for target selection and rational experimental design. New experimental results can then be used to further refine the model. We expect that a rather complete network model can be built after a few iterations.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

This research was supported in part by the US Departmentof Energy's Genomes to Life program (http://doegenomestolife.org/) under project, 'Carbon Sequestration in *Synechococcus sp.*: From Molecular Machines to Hierarchical Modeling' (www.genomes2life.org) and by the National Science

Foundation (NSF/DBI-0354771, NSF/ITR-IIS-0407204). The authors would also like to thank the two anonymous reviewers for their suggestions that largely improve the paper. Funding to pay the Open Access publication charges for this article was provided by the University of Georgia.

Conflict of interest statement. None declared.

REFERENCES

1. Hartwell,L.H., Hopfield,J.J., Leibler,S. and Murray,A.W. (1999) From molecular to modular cell biology. *Nature*, **402**, C47–C52.
2. Oltvai,Z.N. and Barabasi,A.L. (2002) Systems biology. Life's complexity pyramid. *Science*, **298**, 763–764.
3. Kitano,H. (2002) Computational systems biology. *Nature*, **420**, 206–210.
4. Partensky,F., Hess,W.R. and Vaulot,D. (1999) *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. *Microbiol. Mol. Biol. Rev.*, **63**, 106–127.
5. Scanlan,D.J. (2003) Physiological diversity and niche adaptation in marine *Synechococcus*. *Adv. Microb. Physiol.*, **47**, 1–64.
6. Palenik,B., Brahamsha,B., Larimer,F.W., Land,M., Hauser,L., Chain,P., Lamerdin,J., Regala,W., Allen,E.E., McCarren,J. *et al.* (2003) The genome of a motile marine *Synechococcus*. *Nature*, **424**, 1037–1042.
7. Rocap,G., Larimer,F.W., Lamerdin,J., Malfatti,S., Chain,P., Ahlgren,N.A., Arellano,A., Coleman,M., Hauser,L., Hess,W.R. *et al.* (2003) Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature*, **424**, 1042–1047.
8. Dufresne,A., Salanoubat,M., Partensky,F., Artiguenave,F., Axmann,I.M., Barbe,V., Duprat,S., Galperin,M.Y., Koonin,E.V., Le Gall,F. *et al.* (2003) Genome sequence of the cyanobacterium *Prochlorococcus marinus* SS120, a nearly minimal oxyphototrophic genome. *Proc. Natl Acad. Sci. USA*, **100**, 10020–10025.
9. Su,Z., Dam,P., Chen,X., Olman,V., Jiang,T., Palenik,B. and Xu,Y. (2003) Computational inference of regulatory pathways in microbes: an application to phosphorus assimilation pathways in *Synechococcus* sp. WH8102. *Genome Inform. Ser. Workshop Genome Inform*, **14**, 3–13.
10. Dam,P., Su,Z., Olman,V. and Xu,Y. (2004) *In silico* reconstruction of the carbon fixation pathway in *Synechococcus* sp. WH8102. *J. Biol. Sys.*, **12**, 97–125.
11. Capone,D.G. (2000) The marine nitrogen cycle. In Kirchman,D. (ed.), *Microbial Ecology of the Ocean*. Wiley-Liss, NY, pp. 455–493.
12. Flores,E. and Herrero,A. (2005) Nitrogen assimilation and nitrogen control in cyanobacteria. *Biochem. Soc. Trans.*, **33**, 164–167.
13. Reitzer,L. (2003) Nitrogen assimilation and global regulation in *Escherichia coli*. *Annu. Rev. Microbiol.*, **57**, 155–176.
14. Su,Z., Olman,V., Mao,F. and Xu,Y. (2005) Comparative genomics analysis of NtcA regulons in cyanobacteria: regulation of nitrogen assimilation and its coupling to photosynthesis. *Nucleic Acid Res.*, **33**, 5156–5171.
15. Fadi Aldehni,M., Sauer,J., Spielhaupter,C., Schmid,R. and Forchhammer,K. (2003) Signal transduction protein P(II) is required for NtcA-regulated gene expression during nitrogen deprivation in the cyanobacterium *Synechococcus elongatus* strain PCC 7942. *J. Bacteriol.*, **185**, 2582–2591.
16. Sauer,J., Schreiber,U., Schmid,R., Volker,U. and Forchhammer,K. (2001) Nitrogen starvation-induced chlorosis in *Synechococcus* PCC 7942. Low-level photosynthesis as a mechanism of long-term survival. *Plant Physiol.*, **126**, 233–243.
17. Gorl,M., Sauer,J., Baier,T. and Forchhammer,K. (1998) Nitrogen-starvation-induced chlorosis in *Synechococcus* PCC 7942: adaptation to long-term survival. *Microbiology*, **144**, 2449–2458.
18. Sauer,J., Gorl,M. and Forchhammer,K. (1999) Nitrogen starvation in *Synechococcus* PCC 7942: involvement of glutamine synthetase and NtcA in phycobiliprotein degradation and survival. *Arch. Microbiol.*, **172**, 247–255.
19. Chen,X., Su,Z., Dam,P., Palenik,B., Xu,Y. and Jiang,T. (2004) Operon prediction by comparative genomics: an application to the *Synechococcus* sp. WH8102 genome. *Nucleic Acids Res.*, **32**, 2147–2157.
20. Chen,X., Su,Z., Xu,Y. and Jiang,T. (2004) Computational prediction of operons in *Synechococcus* sp. WH8102. *Genome Inform. Ser. Workshop Genome Inform.*, **15**, 211–222.
21. Mao,F., Su,Z., Olman,V., Dam,P., Liu,Z. and Xn,Y. (2006) Mapping of Orthologous Genes in the Context of Biological Pathways: an Application of Integer Programming. *Proc. Natl Acad. Sci. U.S.A.*, **103**, 129–134.
22. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
23. Mushegian,A.R. and Koonin,E.V. (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc. Natl Acad. Sci. USA*, **93**, 10268–10273.
24. Schmidt,H.A., Strimmer,K., Vingron,M. and von Haeseler,A. (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, **18**, 502–504.
25. Uetz,P., Giot,L., Cagney,G., Mansfield,T.A., Judson,R.S., Knight,J.R., Lockshon,D., Narayan,V., Srinivasan,M., Pochart,P. *et al.* (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.
26. Gavin,A.C., Bosche,M., Krause,R., Grandi,P., Marzioch,M., Bauer,A., Schultz,J., Rick,J.M., Michon,A.M., Cruciat,C.M. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.
27. Xenarios,I., Salwinski,L., Duan,X.J., Higney,P., Kim,S.M. and Eisenberg,D. (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.*, **30**, 303–305.
28. Rain,J.C., Selig,L., De Reuse,H., Battaglia,V., Reverdy,C., Simon,S., Lenzen,G., Petel,F., Wojcik,J., Schachter,V. *et al.* (2001) The protein–protein interaction map of *Helicobacter pylori*. *Nature*, **409**, 211–215.
29. Marcotte,E.M., Pellegrini,M., Ng,H.L., Rice,D.W., Yeates,T.O. and Eisenberg,D. (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science*, **285**, 751–753.
30. Enright,A.J., Iliopoulos,I., Kyrpides,N.C. and Ouzounis,C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.
31. Blanchette,M. and Tompa,M. (2002) Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res.*, **12**, 739–748.
32. McGuire,A.M., Hughes,J.D. and Church,G.M. (2000) Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome Res.*, **10**, 744–757.
33. Gelfand,M.S. (1999) Recognition of regulatory sites by genomic comparison. *Res. Microbiol.*, **150**, 755–771.
34. Herrero,A., Muro-Pastor,A.M. and Flores,E. (2001) Nitrogen control in cyanobacteria. *J. Bacteriol.*, **183**, 411–425.
35. Olman,V., Xu,D. and Xu,Y. (2003) CUBIC: identification of regulatory binding sites through data clustering. *J. Bioinform. Comput. Biol.*, **1**, 21–40.
36. Pellegrini,M., Marcotte,E.M., Thompson,M.J., Eisenberg,D. and Yeates,T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.
37. Cormen,T.H., Leiserson,C.E. and Rivest,R.L. (1989) *Introduction to Algorithms*. The MIT Press, Cambridge, Massachusetts.
38. Morel,F.M.M., Rueter,J.G., Anderson,D.M. and Guillard,R.R.L. (1979) Aquil: a chemically defined phytoplankton culture medium for trace metal studies. *J. Phycol.*, **15**, 135–141.
39. Waterbury,J.B. and Willey,J.M. (1988) Isolation and growth of marine planktonic cyanobacteria. *Meth. Enzymol.*, **167**, 100–105.
40. Guillard,R.R.L. (1975) *Culture of Phytoplankton for Feeding Marine Invertebrates*. Plenum, NY.
41. Saeed,A.I., Sharov,V., White,J., Li,J., Liang,W., Bhagabati,N., Braisted,J., Klapa,M., Currier,T., Thiagarajan,M. *et al.* (2003) TM4: a free, open-source system for microarray data management and analysis. *Biotechniques*, **34**, 374–378.
42. Tusher,V.G., Tibshirani,R. and Chu,G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.
43. Frias,J.E., Flores,E. and Herrero,A. (2000) Activation of the *Anabaena* nir operon promoter requires both NtcA (CAP family) and NtcB (LysR family) transcription factors. *Mol. Microbiol.*, **38**, 613–625.
44. Ramasubramanian,T.S., Wei,T.F. and Golden,J.W. (1994) Two *Anabaena* sp. strain PCC 7120 DNA-binding factors interact with vegetative cell- and heterocyst-specific genes. *J. Bacteriol.*, **176**, 1214–1223.

45. Jiang,F., Hellman,U., Sroga,G.E., Bergman,B. and Mannervik,B. (1995) Cloning, sequencing, and regulation of the glutathione reductase gene from the cyanobacterium *Anabaena* PCC 7120. *J. Biol. Chem.*, **270**, 22882–22889.
46. Garcia-Dominguez,M., Reyes,J.C. and Florencio,F.J. (2000) NtcA represses transcription of gifA and gifB, genes that encode inhibitors of glutamine synthetase type I from *Synechocystis* sp. PCC 6803. *Mol. Microbiol.*, **35**, 1192–1201.
47. Harano,Y., Suzuki,I., Maeda,S., Kaneko,T., Tabata,S. and Omata,T. (1997) Identification and nitrogen regulation of the cyanase gene from the cyanobacteria *Synechocystis* sp. strain PCC 6803 and *Synechococcus* sp. strain PCC 7942. *J. Bacteriol.*, **179**, 5744–5750.
48. Bird,C. and Wyman,M. (2003) Nitrate/nitrite assimilation system of the marine picoplanktonic cyanobacterium *Synechococcus* sp. strain WH 8103: effect of nitrogen source and availability on gene expression. *Appl. Environ. Microbiol.*, **69**, 7009–7018.
49. Sakamoto,T., Inoue-Sakamoto,K. and Bryant,D.A. (1999) A novel nitrate/nitrite permease in the marine cyanobacterium *Synechococcus* sp. strain PCC 7002. *J. Bacteriol.*, **181**, 7363–7372.
50. Luque,I., Flores,E. and Herrero,A. (1994) Molecular mechanism for the operation of nitrogen control in cyanobacteria. *EMBO J.*, **13**, 2862–2869.
51. Ramasubramanian,T.S., Wei,T.F., Oldham,A.K. and Golden,J.W. (1996) Transcription of the *Anabaena* sp. strain PCC 7120 ntcA gene: multiple transcripts and NtcA binding. *J. Bacteriol.*, **178**, 922–926.
52. Von Mering,C., Krause,R., Snel,B., Cornell,M., Oliver,S.G., Fields,S. and Bork,P. (2002) Comparative assessment of large-scale datasets of protein–protein interactions. *Nature*, **417**, 399–403.
53. Barabasi,A.L. and Bonabeau,E. (2003) Scale-free networks. *Sci Am.*, **288**, 60–69.
54. Javelle,A., Severi,E., Thornton,J. and Merrick,M. (2004) Ammonium sensing in *Escherichia coli*. Role of the ammonium transporter AmtB and AmtB-GlnK complex formation. *J. Biol. Chem.*, **279**, 8530–8538.
55. Javelle,A. and Merrick,M. (2005) Complex formation between AmtB and GlnK: an ancestral role in prokaryotic nitrogen control. *Biochem. Soc. Trans.*, **33**, 170–172.
56. Muro-Pastor,A.M., Herrero,A. and Flores,E. (2001) Nitrogen-regulated group 2 sigma factor from *Synechocystis* sp. strain PCC 6803 involved in survival under nitrogen stress. *J. Bacteriol.*, **183**, 1090–1095.
57. Peddie,C.J., Cook,G.M. and Morgan,H.W. (1999) Sodium-dependent glutamate uptake by an alkaliphilic, thermophilic *Bacillus* strain, TA2.A1. *J. Bacteriol.*, **181**, 3172–3177.
58. Marsac,N.T. d., Lee,H.M., Hisbergues,M., Castets,A.M. and Bédu,S. (2001) Control of nitrogen and carbon metabolism in cyanobacteria. *J. Appl. Phycol.*, **13**, 287–292.
59. Heinrich,A., Maheswaran,M., Ruppert,U. and Forchhammer,K. (2004) The *Synechococcus elongatus* P signal transduction protein controls arginine synthesis by complex formation with N-acetyl-L-glutamate kinase. *Mol. Microbiol.*, **52**, 1303–1314.