Ancient association of cyanobacterial multicellularity with the regulator HetR and an RGSGR pentapeptide-containing protein (PatX)

Jeff Elhai[1] and Ivan Khudyakov[2]

1. *Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, VA 23284, USA*
2. *All-Russia Research Institute for Agricultural Microbiology, Saint-Petersburg 196608, Russia*

Correspondence to: Jeff Elhai, ElhaiJ@VCU.Edu, (Tel) 1-804-828-0794 (Fax) 1-804-828-0503

Running title: Multicellularity association with HetR and PatX

Key words: cyanobacteria, multicellularity, HetN, PatS, phylogeny, regulation

Summary

One simple model to explain biological pattern postulates the existence of a stationary regulator of differentiation that positively affects its own expression, coupled with a diffusible suppressor of differentiation that inhibits the regulator's expression. The first has been identified in the filamentous, heterocyst-forming cyanobacterium, *Anabaena* PCC 7120 as the transcriptional regulator, HetR, and the second as the small protein, PatS, which contains a critical RGSGR motif that binds to HetR. HetR is present in almost all filamentous cyanobacteria, but only a subset of heterocyst-forming strains carry proteins similar to PatS. We identified a third protein, PatX that also carries the RGSGR motif and is coextensive with HetR. Amino acid sequences of PatX contain two conserved regions: the RGSGR motif and a hydrophobic N-terminus. Within 69 nt upstream from all instances of the gene is a DIF1 motif correlated in *Anabaena* with promoter induction in developing heterocysts, preceded in heterocyst-forming strains by an apparent NtcA-binding site, associated with regulation by nitrogen-status. Consistent with a role in the simple model, PatX is expressed dependent on HetR and acts to inhibit differentiation. The acquisition of the PatX/HetR pair preceded the appearance of both PatS and heterocysts, dating back to the beginnings of multicellularity.

Introduction

In 1952, Alan Turing presented a theoretical mechanism by which homogenous cells could be transformed into patterns of differentiated cells owing to the action of two interacting regulators with different diffusion rates (Turing, 1952). The idea was refined by Gierer and Meinhardt (Gierer and Meinhardt, 1972; Meinhardt, 2008) into the model shown conceptually in Fig. 1A, relying on the actions of a slowly-diffusing, autocatalytic regulator (R) and a rapidly diffusing suppressor molecule (S). While the simplicity of the model is appealing, until recently, there has been little evidence from multicellular eukaryotes to support the theory (Marcon and Sharpe, 2012).

The most compelling biological case for the relevance of the model comes arguably from the multicellular prokaryote, *Anabaena* PCC 7120, which differentiates specialized cells at semi-regular intervals along its filaments (Fig. 2C). These cells, called heterocysts, provide the conditions required for nitrogen-fixation in the presence of molecular oxygen (Kumar *et al.*, 2010; Maldener *et al.*, 2014). *Anabaena* PCC 7120 has been shown to synthesize two proteins, HetR and PatS, that seem to possess the characteristics called for by Gierer and Meinhardt's R and S morphogens. In addition, a third component, NtcA, ties the morphogenetic machinery to the nitrogen status of cells, as is physiologically appropriate, and a fourth, HetN, serves as an S morphogen tailored to pattern maintenance rather than creation. Recent models of heterocyst differentiation have expanded on the ideas of Turing and Gierer and Meinhardt, acknowledging the non-diffusibility of HetR, interactions amongst multiple actors, and their discrete concentrations in a series of cells (Gerdtzen *et al.*, 2009; Herrero *et al.*, 2016; Muñoz-Garcia and Ares, 2016). A prevailing model of the regulation of heterocyst differentiation is summarized in Fig. 1B and fleshed out below.

The gene encoding HetR was identified as part of a hunt for mutants of *Anabaena* PCC 7120 unable to sustain heterocyst differentiation (Buikema and Haselkorn, 1991a), and the protein was soon found to have the characteristics expected from an R morphogen. HetR appears to be a

master regulator of heterocyst differentiation. First, differentiation was abolished by a point mutation in HetR (S179N) and a deletion or disruption of the gene (Buikema and Haselkorn, 1991b; Black *et al.*, 1993). Second, three-fold more heterocysts were formed (including multiple contiguous heterocysts seldom seen in wild-type *Anabaena*) when HetR was expressed from a multicopy plasmid (Buikema and Haselkorn, 1991), from a regulatable promoter (Buikema and Haselkorn, 2001), or from a mutant allele of HetR, R223W (Khudyakov and Golden, 2004). HetR affects the expression of hundreds of genes, with heightened expression after nitrogen deprivation and repression in nitrogen-replete medium (Mitschke *et al.*, 2011; Videau *et al.*, 2014a). Expression of HetR is autocatalytic -- the protein is required for the increase in its own synthesis after nitrogen deprivation (Black *et al.*, 1993; Buikema and Haselkorn, 2001; Cai and Wolk, 1997). HetR acts as a DNA-binding protein (Huang *et al.*, 2004; Flaherty *et al.*, 2014), in a tetrameric state that is regulated by phosphorylation (Valladares *et al.*, 2016), and is in a positive feedback loop with NtcA (Muro-Pastor *et al.*, 2002).

The characteristics of PatS protein is suggestive of its functioning as an S morphogen. PatS was discovered from the ability of a small DNA fragment on a multicopy plasmid to inhibit heterocyst differentiation in *Anabaena* PCC 7120 (Yoon and Golden, 1998). Inhibition required the expression of an 11- to 17-amino acid open reading frame (ORF), called *patS*. Strains lacking this ORF and mutants with ORFs altered in one of the last five codons produced aberrant, though non-random, spacing of heterocysts, including multiple contiguous heterocysts (Yoon and Golden, 1998; Yoon and Golden, 2001). Overexpression of *patS* reduces transcription from an inducible *hetR* promoter (Rajagopalan and Callahan, 2010). Exogenous application of a pentapeptide consisting of the last five amino acids of PatS (RGSGR) blocks heterocyst differentiation (Yoon and Golden, 1998) and promotes posttranslational decay of HetR protein (Risser and Callahan, 2009). This peptide also binds to HetR protein, preventing its binding to DNA (Huang *et al.*, 2004; Feldmann *et al.*, 2012). Maximal binding affinity is achieved by a six-amino acid peptide ending in RGSGR, where the identity of the first amino acid is not critical (Feldmann *et al.*, 2012). With the finding that PatS expression is localized to developing cells (Yoon and Golden, 2001) and is dependent upon HetR (Huang *et al.*, 2004), the matching of PatS characteristics to those of an S morphogen is complete, with one exception: diffusion. Clever experiments have suggested the diffusion of PatS-derived signals to adjacent cells (Risser and Callahan, 2009; Corrales-Guerrero *et al.*, 2013; Rivers *et al.*, 2014), but direct evidence for diffusion of any morphogen in *Anabaena* PCC 7120 remains elusive.

A second putative S morphogen was recognized within the protein HetN, initially identified in a similar fashion as PatS: (1) its presence on a multicopy plasmid suppressed heterocyst differentiation in *Anabaena* PCC 7120, and (2) its interruption led to multiple contiguous heterocysts (Black and Wolk, 1994; Bauer *et al.*, 1997). Although *hetN* encodes a protein similar to short chain dehydrogenases (Pfam PF00106) and polyketide synthases (Pfam PF008659), most of the protein could be deleted without affecting its ability to suppress heterocyst differentiation (Higa *et al.*, 2012). However, an RGSGR sequence found within HetN (Li *et al.*, 2002) proved to be essential for the protein's effectiveness as a suppressor (Higa *et al.*, 2012; Corrales-Guerrero *et al.*, 2014). HetN differs from PatS in two important respects. First, its expression starts at a late stage of differentiation and persists in mature heterocysts (as opposed to induction at an early stage and downregulation in mature heterocysts) (Callahan and Buikema, 2001; Videau *et al.*, 2014b). Second, HetN is important in the maintenance of the pattern of heterocysts but not its initial formation (Callahan and Buikema, 2001).

The functions of HetR, PatS, and HetN in *Anabaena* PCC 7120, combined with the Turing/Meinhardt model for pattern formation (Fig. 1B), provide an appealing explanation for the appearance of spaced heterocysts in response to nitrogen deprivation. However, there are several observations that do not obviously square with this view. HetN-like proteins bearing the RGSGR motif are found in only a small fraction of heterocyst-forming cyanobacteria (Corrales-Guerrero *et al.*, 2014). ORFs capable of producing a PatS-like protein were also reported to be absent in heterocyst-forming *Cylindrospermopsis raciborskii* CS-505 (Stucken *et al.*, 2010) and *Anabaena* 90 (Wang *et al.*, 2012). Finally, proteins antigenically similar to HetR are found even in filamentous cyanobacteria that don't make heterocysts (Zhang *et al.*, 2009), consistent with the presence in such strains (henceforth termed "non-het-filamentous") of DNA that hybridized to a hetR probe (Buikema and Haselkorn, 1991). Cyanobacteria without the postulated machinery evidently produce spaced heterocysts, and cyanobacteria with at least part of the machinery do not. These observations prompted us to look systematically in cyanobacterial genomes for genes that may encode the proteins that make up the machinery of the Turing/Meinhardt model.

Results

Phylogeny of cyanobacteria used in this study

In order to place the presence or absence of NtcA, HetR, PatS, and HetN in a logical context, we developed a phylogenetic tree of the 127 core cyanobacterial genomes used in this study (see Supporting Table S1 for description of the genomes). The tree is based on alignments of 29 proteins, a subset of the 32 proteins used by Howard-Azzeh et al (2014). Trees based on 16S rRNA sequences provide significantly less information on which to base a tree, and consequently there are far fewer nodes with good bootstrap information than with trees based on many conserved proteins (Shih *et al.*, 2014). Not surprisingly, considering its basis, the tree is completely concordant with that of Howard-Azzeh et al., with respect to the 100 genomes used by both groups and nodes enjoying strong bootstrap support. It is also concordant with three other trees based on different sets of concatenated protein alignments (Sánchez-Baracaldo *et al.*, 2014; Shih *et al.*, 2014; Schirrmeister *et al.*, 2015). All of these trees differ in important respects from that of Uyeda et al (2016), who attempted to avoid artifacts resulting from long branch attraction. Their tree differs in the placement of the picocyanobacteria, *Prochlorothrix hollandica* PCC 9006, branching heterocyst-forming cyanobacteria, and the *Calothrix* PCC 6303 and PCC 7103 pair. However, none of these differences in predicted phylogeny affect the arguments we will present. The intermixing of branched and non-branched (Sections IV and V) heterocyst-forming cyanobacteria is discussed later.

Fig. 2 shows one of many possible interpretations of the cyanobacterial phylogenetic tree. The tree is rooted by *Gloeobacter violaceus* PCC 7421, owing to the early divergence of *Gloeobacter* from the rest of the cyanobacterial lineage (Saw *et al.*, 2013; Schirrmeister *et al.*, 2015). If that rooting is accurate, then it is evident that unicellularity is the original morphotype of cyanobacteria. Multicellular filamentous strains appear to have arisen early and include in their number the coherent clade of heterocyst-forming cyanobacteria. Whether multicellularity appeared once as shown in Fig. 2 or multiple times is an open question, one that is discussed later. In this article, we use "filamentous" to refer to a specific morphology and "multicellularity" to refer to a life style that implies functional interactions between cells (Schirrmeister *et al.*, 2013; Herrero *et al.*, 2016).

Fig. 3 shows details of the phylogenetic tree, split between heterocyst-forming cyanobacteria (Fig. 3A) and the rest (Fig. 3B), and also lists the genome abbreviations used in this work.

Appearance of HetR in cyanobacteria

Orthologs of HetR were found in almost all filamentous strains and in almost no unicellular strains (Fig. 3). The filamentous exceptions are the *Pseudanabaenas* (Clade 8 in Figs. 3 and 4B) and *Geitlerinema* PCC 7105. The absence of HetR in *Geitlerinema* PCC 7105 may be the result of an incomplete sequence or mis-assembly of the genome. The gene order surrounding *hetR* (Supporting Fig. S1) is conserved in the two closest available genomes, those of *Phormidium* OSCR and *Phormidium* BDU 130791. Several genes near *hetR* in the two *Phormidia* are completely missing from *Geitlerinema* PCC 7105, including five ribosomal proteins presumably required for life that are found in all other cyanobacterial genomes (except in one case where the gene cluster is split between two contigs). We conclude that a segment containing these genes as well as *hetR* is almost certainly present in *Geitlerinema* PCC 7105 but missing from its available genome assembly.

The only phenotypically unicellular strains possessing a HetR ortholog are *Synechococcus* PCC 7002 and *Synechococcus* PCC 7335. *Synechococcus* PCC 7002 is closely related to *Leptolyngbya* PCC 7376, a filamentous strain. It was formerly called *Agmenellum quadruplicatum* PR6 (Rippka *et al.*, 1979) because of its propensity to grow as four-cell filaments, and a variant has been found that forms long filaments at 24^{o} C (Don Bryant, personal communication). *Synechococcus* PCC 7335 lies phylogenetically within a clade otherwise consisting of filamentous cyanobacteria (Fig. 3B) and is most closely related to *Leptolyngbya Heron Island J*. HetR may therefore be a holdover from a time in the recent evolutionary past when the ancestors of these two strains were filamentous.

The tight association of HetR with filamentous strains raises the possibility that the protein may be important in the multicellular life style (at least in the clade that excludes the *Pseudanabaena*). If so, then one might expect the phylogeny of the HetR protein to match the organismal phylogeny, if multicellularity arose only once, but not if multicellularity arose several times. In fact, though the HetR phylogenetic tree (Supporting Fig. S2) lacks sufficient bootstrap support to be definitive, it matches the phylogenetic tree as well as can be expected. In particular, HetRs from Clade 1A (Figs. 3 and 4A), containing all the heterocyst-forming cyanobacteria, appear to have a common ancestor. HetRs from *Synechococcus* PCC 7002, *Leptolyngbya* PCC 7376, and *Spirulina subsalsa* PCC 9445, all in Clade 2 (Figs. 3 and 4B), form a coherent group distinct from other HetR proteins., all with bootstrap support and consistent with the phylogenetic tree. It is also worth noting that there are seven organisms represented in Supporting Fig. S2 with more than one apparent copy of HetR. In each case, there is one copy of HetR (termed "primary") that has a typical amino acid sequence (see below), while the other copies (termed "secondary") have less conserved sequences and cluster together (Supporting Fig. S2).

While the evolutionary connection of all the HetR sequences is beyond dispute, it is an open question as to whether the HetR proteins in the phenotypically unicellular strains and the secondary HetR proteins have the same function as primary HetR proteins in the filamentous strains or indeed any function at all. To address this question, all available HetR sequences were aligned (Fig. 4 and Supporting Fig. S3). There is overwhelming amino acid sequence conservation in the 75 primary HetR proteins from filamentous cyanobacteria (including

heterocyst-forming). Of the 299 amino acid positions (allowing for frayed N- and C-termini), 172 (the green columns) are highly conserved as defined in Fig. 4. Of this latter group, 23 residues have been implicated in DNA- (19) or PatS-binding (4), from analyses of crystal structures (Kim *et al.*, 2011; Kim *et al.*, 2013; Hu *et al.*, 2015) and in vitro assays (Risser and Callahan, 2007; Kim *et al.*, 2011; Feldmann *et al.*, 2012; Kim *et al.*, 2013; Hu *et al.*, 2015) and in vivo phenotypes of site-specific mutants (Buikema and Haselkorn, 1991; Dong *et al.*, 2000; Huang *et al.*, 2004; Khudyakov and Golden, 2004; Risser and Callahan, 2007; Feldmann *et al.*, 2011; Kim *et al.*, 2013; Hu *et al.*, 2015). Only four primary HetR sequences in filamentous cyanobacteria have mutations in any residue implicated in DNA- or PatS-binding, and three of the mutations are conservative.

The two phenotypically unicellular strains present a different picture. To avoid observation bias (there are far more available sequences of *Nostocs* and *Anabaenas* than sequences of strains phylogenetically close to*Synechoccocus* PCC 7002 and PCC 7335), we compared each of the two strains to its closest relative: *Synechococcus* PCC 7335 to *Leptolyngbya* Heron Island J and *Synechococcus* PCC 7002 to *Leptolyngbya* PCC 7376 (Table 1). *Synechococcus* PCC 7335 has more than four times the number of mutations in conserved positions as does *Leptolyngbya* Heron Island J, and 24% are non-conservative substitutions, compared to 0% for *Leptolyngbya* Heron Island J. With the more distant *Synechococcus* PCC 7002 / *Leptolyngbya* PCC 7376 pair, the former has 2.6-times more mutations than the latter and 62% non-conservative substitutions, compared to 42% for *Leptolyngbya* PCC 7376. HetRs from the unicellular strains are evidently under lesser or different selection than those from the related filamentous strains. Similarly, secondary HetRs have a much higher number of deviations and non-conservative deviations than their primary counterparts (Table 1 and data not shown).

If the high number of deviations in unicellular and secondary HetRs were due to drift in the absence of selection, then one would expect to find deviations spread randomly across the functional categories, but this is not observed (Table 1). In both cases, the amino acids implicated in DNA-binding are significantly less likely to deviate from the standard residue. Two secondary HetRs, $HetR_{Lep6406\text{-}b}$ and $HetR_{Lyn141951\text{-}b}$, have the number of deviations expected by chance, but the other secondary HetRs have far fewer (data not shown). In contrast, the HetRs from unicellular cyanobacteria are significantly more likely to experience deviations in residues associated with PatS-binding. In addition, HetR from *Synechococcus* PCC 7002 also carries an R223A mutation, in a residue implicated in the phenotypic sensitivity of HetR to PatS and HetN (Khudyakov and Golden, 2004). Evidently, mutations are not random in secondary HetRs and those in unicellular cyanobacteria, indicating maintained selective pressure during at least part of the period since separating from their primary filamentous homologues, presumably owing to retained DNA-binding function.

Appearance of HetN and PatS in cyanobacteria

If HetN is defined as a protein (a) similar in sequence to HetN of *Anabaena* PCC 7120 and (b) possessing RGSGR, then its incidence is limited to *Anabaena* PCC 7120 and its three closest relatives (Fig. 3A), plus two distantly related unicellular cyanobacteria lacking HetR (Fig. 3B). In addition, two strains of *Chlorogloeopsis* carry a HetN-like protein with the sequence ERGSGH, one off from the conventional motif (Fig.4A and Supporting Fig. S4). There is good reason to doubt the significance of these proteins in heterocyst regulation, as a mutation of the *Anabaena* PCC 7120 HetN motif from RGSGR to RGSGK results in loss of function in *Anabaena* PCC 7120 (Higa *et al.*, 2012). A phylogenetic analysis of HetN-like proteins

(Supporting Fig. S4) shows a well-supported cluster of *Anabaena* PCC 7120 HetN and its three relatives, lying distinct from a second well supported cluster that includes the two *Chlorogloeopsis* HetN candidates. The two unicellular HetN candidates lie in a distant cluster (not shown).

It is much more difficult to identify putative PatS proteins. Only three genomes amongst the 127 core genomes considered in this work have had *patS* genes annotated within them: *Anabaena* PCC 7120 (Yoon and Golden, 1998), *Nostoc punctiforme* ATCC 29133 (Meeks *et al.*, 2002), and *Nodularia spumigena* CCY9414 (Voβ *et al.*, 2013). Two other genomes, *Leptolyngbya* NIES 3755 and *Arthrospira* PCC 8005, have genes misannotated as *patS*. The lack of annotated *patS* genes is to be expected, since most automated gene-calling processes exclude ORFs of the size of *patS*. Scanning genomes for ORFs containing RGSGR is also unsatisfactory, as the rate of false positives is far too high. The genomes considered in this study average 11.1 RGSGR-containing ORFs, of which 82% are in called genes (80% of these are in a conflicting reading frame). The high-GC genome of *Cyanobium gracile* PCC 6307 provides an extreme example: it has 68 RGSGR-containing ORFs. It is highly unlikely that any of them have regulatory function in this unicellular organism. To identify true orthologs of PatS, we therefore also considered genetic context.

The *patS* gene from *Anabaena* PCC 7120 (*asl2301*) is preceded by two genes (*all2302* and *all2303*) encoding proteins annotated as patatin and dihydroorotase, respectively. On the downstream side is a gene (*alr2300*) annotated as *hetY*, encoding a protein described as necessary for timely heterocyst differentiation (Yoon *et al.*, 2003). We found 28 genomes, all from heterocyst-forming cyanobacteria, that have a short RGSGR-containing ORF situated near at least one of the typical upstream or downstream genes (Fig. 5 and Supporting Table S4). The orientation of the ORF relative to the neighbor gene(s) (parallel upstream, parallel downstream, convergent, or divergent) is in all cases the same as that of the analogous *patX*-gene pair from *Anabaena* PCC 7120,

To assess whether the remaining 11 heterocyst-forming cyanobacteria possess *patS*-like ORFs that either lack a complete RGSGR motif or reside in non-canonical genetic locations, we used the sequence characteristics of the 28 putative PatS proteins as the basis of a systematic search. The genomes of each heterocyst-forming cyanobacterium was scanned in all six reading frames with a position-specific scoring matrix (PSSM), applying the aggregated positional amino acid frequencies at the RGSGR motif and 8 amino acids upstream (see Methods for details). Each of the 20 to 80 million virtually translated ORF fragments generated from a genome yielded a score representing the likelihood of the fragment arising from the positional frequencies of the PatS PSSM as compared to chance. This method was able to pick out the contextually determined *patS* ORFS as having the top score of all fragments, usually by orders of magnitude (Supporting Fig. S5(A) and Table S8). PatS from the symbiotic strain *Richellia intracellularis* HH01 is exceptional in that its score is so low as to be mixed in with the right tail of the mass of random scores (Supporting Fig. S5B).

Of particular interest were the scores of ORF fragments from genomes without already recognized PatS candidates. In two cases (*Calothrix* PCC 6303 and *Scytonema tolypothrichoides* VB61278), a scan of the genome picked out a plausible PatS sequence (Fig. 5, Supporting Fig. S5(B), and Supporting Table S4). The ORF fragment of *Cylindrospermum stagnali* PCC 7417 with the highest score is less compelling. The remaining 8 strains without PatS

candidates fall in the clade shared with *Anabaena cylindrica* PCC 7122. No ORF fragment in these genomes have high scores or otherwise look plausible.

A glutamate residue (E) just before the RGSGR motif was previously noted in both PatS and HetN from *Anabaena* PCC 7120 (Corrales-Guerrero *et al.*, 2014), and it is nearly universal amongst the candidate PatS proteins (Fig. 5), substituted only by aspartate (D). The position of the motif within 10 amino acids of the N-terminus appears also to be general, with the protein from *Rivularis* as the sole exception. However, the C-terminal position of RGSGR in PatS is not conserved -- 43% of the candidate proteins have C-terminal extensions, particularly common in the *Scytonema*/*Tolypothrix* clade.

Appearance of PatX in cyanobacteria

The absence of both HetN and PatS in the *Anabaena cylindrica* PCC 7122 clade despite normal heterocyst spacing could be explained if these genomes possess a third RGSGR-containing protein. We could identify only one protein containing RGSGR that can be found in the genomes of multiple cyanobacteria, including members of the *Anabaena cylindrica* PCC 7122 clade. That protein, termed PatX, is poorly conserved in overall sequence, and so we turned again to genetic context to guide discovery of other members of the family.

Genes encoding RGSGR-bearing proteins were found in the genomes of 33 heterocyst-forming cyanobacteria (Fig. 6A and Supporting Table S5), near at least one of six linked genes that include three known to be related to heterocyst differentiation or function: *hetR*, *sepJ* (also known as *fraG*, encoding a protein required for filament integrity under N-fixing conditions (Nayar *et al.*, 2007)), and *glnA* (encoding glutamine synthetase (Tumer *et al.*, 1983), which catalyzes the first step in the assimilation of fixed nitrogen (Flores and Herrero, 1994)). In *Anabaena* PCC 7120, these genes are *alr2339*, *all2338*, and *alr2328*, respectively. Immediately upstream of the *patX* gene is a gene encoding a protein that is highly conserved in Groups 1-6 (described in Figs. 2 and 3) and that may be an FAD-dependent oxidoreductase (All2333 in *Anabaena* PCC 7120). Amongst the 33 proteins are the two postulated by Stucken et al (2010) to substitute for PatS (Stucken *et al.*, 2010) and another misidentified as PatS (Zhang *et al.*, 2009). Using the same criteria, PatX candidates were found in 21 non-het-filamentous strains and the unicellular *Synechococcus* PCC 7335.

From this collection of proteins, certain structural generalities stand out (Fig. 6A and Fig. 7B) and may be contrasted with those of PatS (Fig. 5 and Fig. 7A). The RGSGR motif lies close to the C-terminus of the protein and is usually preceded by a H or Y (heterocyst-forming strains) or H, E, or D (non-het-filamentous strains) and followed by R. The motif is often preceded by a proline-rich region. The N-termini are rich in hydrophobic residues, and all filamentous organisms except *Microcoleus* PCC 7113 have at least one candidate PatX protein with a signal peptide identified by SignalP (see Experimental Procedures). The N-terminus of heterocyst-forming strains exhibits a striking pattern, *PxxxPPxxxPPxxx*, where *P* is a polar residue, *S*, *T*, or *G*, and *x* is any hydrophobic residue. This motif is found in proteins with one transmembrane domain that form homodimeric complexes (Dawson *et al.*, 2002). However, there is no good evidence that the region forms a transmembrane domain. TMHMM predicts such regions in 18% of the PatX sequences, but the program is often confused by signal sequences (Krogh *et al.*, 2001). Antonaru and Nürnberg (2017) recently recognized proteins they called alternative PatS that share the characteristics of PatX.

The same PSSM-based approach described above regarding PatS was used to find additional PatX candidates, except that two PSSMs were used – one trained on the RGSGR-motif region and the other on the N-terminus. The N-terminal PSSM divided the $1.8x10^9$ ORF fragments from heterocyst-forming strains cleanly into two groups: 37 fragments scoring higher than 7.8 (almost 8 orders of magnitude above chance) and the rest, scoring below 6.5. The first group consisted solely of PatX candidates found by genetic context, plus three new PatX candidates found in *Anabaena* PCC 7120 (i.e. Asl2332) and its close relatives, despite their possession of RGTGR in place of RGSGR (Fig. 6A and Supporting Fig. S5D and F). The RGSGR-region PSSM was almost as discriminatory, finding all but two genetically determined PatX candidates with scores above 6.5 and 11 other ORF fragments. Of these, 3 were the *Anabaena* candidates found earlier, 1 was a PatX-like protein from *Nostoc punctiforme* (in addition to its primary PatX). The remainder lacked the characteristics of PatX. In addition, a potential protein was found in *Richelia intracellularis* HH01 that is only weakly similar to PatX, lacking proline residues upstream from its RGSGR motif and lacking an N-terminal *PxxxPPxxxPPxxx* motif. The two *Chlorogloeopsis* strains show no sign of PatX (see Methods). Apart from them and possibly *Richelia*, all heterocyst forming strains evidently have a PatX protein.

PatX is equally well represented in the genomes of non-het-filamentous strains. The set of 22 candidates identified by the presence of RGSGR and proximity of the genes to *hetR* and/or an *all2333* ortholog (Fig. 7B) was extended as before by motif-specific and N-terminal specific PSSMs trained against the non-het-filamentous set (Fig. 6B and Supporting Table S5). The diversity of sequences comprising the training set relative to the training set from heterocyst-forming cyanobacteria produced a tool that is less discriminating, but nonetheless identified candidate PatX proteins in the remaining 7 non-het-filamentous strains. Five of these candidate PatX proteins are in the appropriate genetic context but have noncanonical motifs: RGTGR (*Crinalium epipsammum* PCC 9333) and RGGGR (four *Planktothrix* strains). The PatX candidates from non-heterocyst-forming strains also have N-termini identified by SignalP as signal sequences, but the *PxxxPPxxxPPxxx* motif is less pronounced, and prolines are less prominent in the region preceding the RGSGR motif. The amino acids immediately preceding and following the RGSGR motif in general follow the tendencies of those in PatX sequences from heterocyst-forming strains (Fig. 7B and 7C). Remarkably, five of the PatX candidates have one to three extra RGSGR motifs at spaced intervals. A byproduct of the analysis was the discovery of 15 PatX-like candidates (termed alternative PatXs) distinct from the primary candidates (Fig. 6B and Supporting Table S5). Of these, 9 were found by both of the PSSMs.

Apart from *Synechococcus* PCC 7335, no unicellular cyanobacterium has an identifiable PatX ORF, despite the fact that one of the genes (an FAD-dependent oxidoreductase) typically found near *patX* orthologs is common amongst unicellular strains and another (*glnA*) is ubitquitous. Of particular note, nothing similar to *patX* could be found in *Synechococcus* PCC 7002 and its siblings (Supporting Table S1), despite a thorough search of nearby sequences.

Sequences upstream from PatS and PatX genes

It might be expected that two proteins, PatS and PatX, that share the RGSGR motif that interacts with HetR might also share transcriptional regulatory motifs. The sequences upstream from the corresponding genes were therefore examined, to consider whether shared elements might serve as the basis for their common regulation. Mitschke et al. (2011) reported two transcriptional start sites upstream from $patS_{Ana7120}$, one at -580 that is induced by nitrogen deprivation and another

at -692 that is not (these differ from the two 5' ends reported earlier (Yoon and Golden, 2001)). The inducible start site is preceded by a consensus DIF1 motif (TCCGGA) (called "DIF+" by Mitschke *et al.*, 2011), centered at -35 (Supporting Fig. S6). Promoters with the DIF1 motif in the -35 region are induced early in heterocyst differentiation and are active in developing heterocysts (Yoon and Golden, 2001; Mitschke et al. 2011; Muro-Pastor, 2014; Muro-Pastor et al., 2017).

In order to learn what characteristics to look for in a functional DIF1 motif, we turned to a collection of such motifs (with no more than one mismatch) that has been shown to precede 58 inducible genes in *Anabaena* PCC 7120, found upstream from DIF1 transcriptional start sites (defined by Mitschke et al. (2011) as those induced at least 8-fold by nitrogen starvation and relying on a functional HetR protein for induction). The DIF1 motifs begin 33 to 38 nucleotides before the transcriptional start site or (three outliers) at -43 or -44. Considering just the 55 regions of the first group, the DIF1 motif is followed 17 or 18 nucleotides downstream by a less conserved motif (G[T/A]ANA]) around 10 nucleotides before the transcriptional start site (Supporting Fig. S6 and Fig. 8D). One might surmise from previous studies (Mitschke *et al.*, 2011; Li *et al.*, 2015) that DIF1 motifs are followed by classical -10 regions recognized by SigA (consensus TATAAT), however the similarity of the -10 region to TATAAT is low, with a median score (L_2(-10) in Supporting Fig. S6) of 0.55 compared to a median score of 4.30 over all *Anabaena* transcriptional sites (Mitschke *et al.*, 2011). Since the scale is based on $\log_2$, the two median scores differ by a factor of 13. There is no obvious correlation between on one hand either the -10 scores, the quality of the DIF1 motif, or the similarity of the -10 region to G[T/A]ANA and on the other hand the number of transcripts at 8h after N-deprivation ($Reads_{8h}$) or degree of induction (Log_2(8h/0h)).

The 55 promoter regions with DIF1 motifs were used as a training set to construct a position specific scoring matrix (see Methods) to identify possible DIF1-motif-containing regions (DIF1 regions) in sequences upstream from *patS* and *patX*, those better predicted by the sequence characteristics of the training set than by the overall nucleotide frequencies of the training set. This strategy led to the identification of candidate regions (with a $\log_{10}$ odds score better than 4) in 15 of the 30 putative *patS* genes (Supporting Fig. S7). Using these 15 regions as the training set led to the discovery of an additional putative DIF1 region. Regions identified in this manner clustered consistent with the phylogenetic relationships shown in Fig. 3. In all cases, the DIF1 motif lies in an intergenic region contiguous with the beginning of the putative *patS* gene. 88% of the regions carry exact matches to the TCCGGA motif. In contrast, only 26% of Mitschke et al's set have exact matches. The consensus -10 region of the putative DIF1 region upstream of *patS* genes is similar to that of the training set of DIF1 regions (compare Fig. 8A with 8D): GTAGAGA vs G[T/A]ANA.

It must be stressed that the mere presence of a DIF1 motif defined as no more than one nucleotide off from TCCGGA is not significant without additional sequence or positional information. Such sequences are found by actual count on average once every 106 to 218 nt over the range of heterocyst-forming cyanobacteria.

The same training set of 55 DIF1 regions was used to identify putative DIF1 regions upstream from *patX* genes. Candidate regions with good scores were found upstream from *patX* genes in all 37 of the heterocyst-forming cyanobacteria that have *patX* (Supporting Fig. S8A). All DIF1 motifs were positioned 57nt upstream from the translational start site (except for one case where it is 58nt), and all were perfect TCCGGA sequences. The regions had a large number of

conserved sequences (Fig. 8B), including a conserved -10 region (always 18 nt from DIF1), GTAnnAG, preceded by a conserved A.

Similarly, each of the 28 non-heterocyst-forming cyanobacteria showed plausible DIF1 regions close to the translational start site of *patX* (Supporting Fig. S8B). Although there are far fewer well conserved positions in the upstream sequences (to be expected, given the greater phylogenetic range of this cyanobacterial grouping – see Fig. 3), there is a conserved cluster of nucleotides near the -10 position, very similar to those of heterocyst-forming cyanobacteria (Fig. 8C). In the seven cases where cyanobacteria have two copies of *patX*, six bear DIF1 regions with the same characteristics as those of other non-heterocyst-forming cyanobacteria. The seventh case, *Prochlorococcus hollandica* PCC 9006, may have a degenerate form. The *patX* genes of two strains of Lyngbya have two DIF1 motifs one after the other preceding the translational start site. Considering all 37 DIF1 motifs, only 11% have perfect TCCGGA sequences, while 65% have TCCTGA, spread over the full range of non-heterocyst-forming strains. The *patX* gene of *Synechococcus* PCC 7335, the only unicellular strain possessing one, is preceded by a respectable DIF1 region. It is important to note that the 11 PatX candidates and 18 alternative PatX candidates identified by the PSSM all have DIF1 motifs properly positioned relative to their start codons, which may serve as a partial confirmation of the efficacy of the method that found them.

A striking feature of sequences upstream from *patX* genes is the presence of NtcA-binding sites ($GTAN_8TAC$) (Picossi *et al.*, 2014) in 33 of 37 heterocyst-forming cyanobacteria, always 13-16nt from the DIF1 motif (Supporting Fig. S8A). The site preceding *patX* from *Anabaena* PCC 7120 has been shown experimentally to bind NtcA (Picossi *et al.*, 2014). No such sites are found upstream from the DIF1 motifs of non-heterocyst-forming cyanobacteria nor from those upstream from *patS* genes in heterocyst-forming strains (Supporting Figs. S7 and S8B). Only two genes of the 57 in with DIF1 transcriptional start sites (Mitschke et al., 2011) have DIF1 motifs preceded by NtcA-binding sites (15nt before the DIF1 motif of *all0935* and 22nt before the DIF1 motif of *asr1775*). The three primary PatX sites that lack plausible NtcA-binding sites are all exceptional: *Richelia intracellularis* HH01, a symbiotic strain that has filaments of only a few cells (Janson *et al.*, 1999); *Cylindrospermopsis raciborski* CS-505, a strain that makes heterocysts only on the termini of its filaments; and *Rhapidiopsis brooki* D9, a related strain that does not complete differentiation to mature heterocysts (Stucken *et al.*, 2010). However, *patX* from a second strain with terminal heterocysts, *Cylindrospermum stagnali* PCC 7417 is preceded by an NtcA-binding site.

Localization of patX transcription and the effect of its overexpression on differentiation.

As shown above, the *patX* promoter region is similar to the previously described DIF1 region associated with genes induced at an early stage of differentiation specifically in prospective heterocysts. To determine whether this is true also in the case of *patX*, we constructed a plasmid, pRIAM971, that can replicate in *Anabaena* PCC 7120 and carries a P_{patX}-*gfp* transcriptional fusion. GFP fluorescence in wild-type *Anabaena* was localized in a distinct pattern as early as 6-8 h after nitrogen step-down (Fig. 9). This occurred well in advance of the gradual decrease in red autofluorescence that accompanies heterocyst differentiation due to the degradation of light-harvesting phycobiliproteins (Toyoshima *et al.*, 2010) and of any morphological changes that manifest cell fate determination. Later on the reporter activity was localized in developing heterocysts. Both the kinetics and pattern of fluorescence induction after combined nitrogen

depletion and the intensity of fluorescence in proheterocysts were similar in strains with P_{patX}-*gfp* and P_{patS}-*gfp* (carried on pAM830; Yoon and Golden, 1998).

Overexpression of *patS* and *hetN* abolish heterocyst differentiation in *Anabaena* PCC 7120 (Rajagopalan and Callahan, 2010; Higa *et al.*, 2012), and so it was of interest to determine whether overexpression of *patX* would have the same effect, despite PatX's S→T substitution in the RGSGR motif. PatX was expressed in *Anabaena* PCC 7120 on a multicopy plasmid, pRIAM810 from a copper-regulated *petE* promoter. Incubation of this strain in liquid or on solid $BG11_0$ medium containing copper completely blocked heterocyst differentiation and diazotrophic growth of the wild type (Fig. 10), indicating that even with the S→T replacement in the RGTGR motif, PatX is capable of suppressing HetR activity and preventing heterocyst differentiation.

Discussion

The Turing/Meinhardt model, expanded by later models (Herrero *et al.*, 2016) calls for an R morphogen that causes developmental action and for a diffusible S morphogen that inhibits it. HetR and PatS, modulated by the actions of NtcA and HetN, may be a realization of this model, leading to the appearance of spaced heterocysts in *Anabaena* PCC 7120. However, PatS is not present in a clade containing many heterocyst-forming cyanobacteria, and HetN is absent from all but a few. Their regulatory burden may be taken up by a third RGSGR-bearing protein, PatX, one that is nearly universal amongst heterocyst-forming cyanobacteria. PatX is expressed in a fashion similar to PatS and its overexpression represses heterocyst differentiation, consistent with an interaction with HetR. It must be pointed out, however, that overexpression of ORFs surely unrelated to heterocyst regulation but nonetheless bearing the RGSGR motif also repress heterocyst differentiation (Wu *et al.*, 2004). We will present elsewhere evidence that in a mutant of Anabaena PCC 7120 in which patS and hetN are knocked out or not expressed, PatX is required to prevent rapid synchronous heterocyst differentiation in the absence of nitrate and ectopic necridia formation in its presence (I. Khudyakov, unpublished). Surprisingly, HetR and PatX are also nearly universal amongst filamentous cyanobacteria that don't make heterocysts and are nearly absent from unicellular strains (and the exceptions are instructive). What need do non-het-filamentous strains have for a Turing/Meinhardt mechanism? Cyanobacteria were among the first multicellular organisms on Earth (Schirrmeister *et al.*, 2015; Herrero *et al.*, 2016). To understand how this innovation arose and the benefit it provided, it would be reasonable to examine those proteins that filamentous cyanobacteria hold dear and that unicellular cyanobacteria count as of no selective value.

Conclusions regarding the prevalence of PatS and PatX rely on the ability to recognize these proteins, but this is difficult by the usual automated methods applied to genomes. PatS poses a formidable challenge for automated methods because of its small size (median size 13 amino acids, Fig. 5). As a result, while *patS* genes was at one time annotated in the two genomes found at NCBI (National Center for Biotechnology Information) with physiological evidence for the function of PatS (*Anabaena* PCC 7120 and *Nostoc punctiforme* ATCC 29133) plus one more (*Nodularia spumigena* CCY9414) without such evidence, reannotation efforts by the NCBI (National Center for Biotechnology Information, 2017) that paid no heed to published results have discarded these annotations. The genes can be recognized, however, by the presence of an

encoded RGSGR motif in a standard genetic context (Fig. 5) and by a genomic scan using positional amino acid frequencies of a training set, as discussed above.

Genes encoding PatX generally appear in recently annotated genomes, but the low level of sequence conservation makes it difficult to cluster them into a family of orthologs. However, they are often readily identified by a combination of genetic context and sequence characteristics: a generally C-terminal RGSGR motif, an N-terminal region, either membrane-spanning or a signal sequence, and both separated by a spacer region generally rich in prolines (Fig. 6). We imagine that the protein might be directed by a signal sequence out of an immature heterocyst and into the periplasm where it is acted on by a periplasmic peptidase to produce a diffusible inhibitor (Yoon and Golden, 1998) or, alternatively, tethered to an internal membrane near pores at a septal junction to increase the local concentration of active peptide resulting from proteolytic processing. The proline-rich region might maintain PatX in a disordered structure to ensure that the RGSGR region is available to a peptidase.

Eight PatX candidates and one alternative PatX candidate have a core motif of RGTGR or RGGGR. It has been reported that mutating the central S in the RGSGR motif to A substantially reduced the regulatory activity of HetN (Higa *et al.*, 2012) and abolished the activity of PatS (Corrales-Guerrero *et al.*, 2013), but conceivably mutations to either T or G would have less effect. Our results demonstrate that the PatX of *Anabaena* PCC 7120 (bearing RGTGR) has significant activity, at least when overexpressed. PatX of *Arthrospira platensis* NIES 39 (Zhang *et al.*, 2009) and *Mastigocladus laminosis* (Antenaru and Nürnberg, 2017), both bearing RGSGR, also suppressed heterocyst differentiation when expressed in *Anabaena* PCC 7120 on a multicopy plasmid.

The ability of the methods used in this work to find such deviant PatX proteins strengthen the case that strains showing no candidate PatX proteins really have none, not even one that lacks one of the five conserved residues. These strains may well have a protein so deviant that it escapes detection, but it would need to be different in multiple respects, both in the RGSGR region and the N-terminus.

Having defined PatS and PatX in this way, we can educe the following generalities. Most importantly, PatX and HetR are present together in almost all filamentous cyanobacteria except for the distantly related *Pseudanabaena* (where both are lacking) and absent in almost all unicellular cyanobacteria (Fig. 3). A clear implication of this finding is that two putative cogs in a Turing/Meinhardt pattern-generating machine have roles beyond heterocyst differentiation, a function under strong selection in filamentous strains but not unicellular strains. HetR from the two exceptional unicellular cyanobacteria, *Synechococcus* PCC 7002 and *Synechococcus* PCC 7335, both are atypical, with many differences relative to closely related filamentous strains and defects in conserved residues associated with the binding of RGSGR (Fig. 4).

These atypical HetR proteins may be nonfunctional (in a state of decay), may retain HetR-like function, or may have transitioned to a function different from canonical HetR. The latter is most likely, as an analysis of mutations indicates continued selection for binding to DNA but not to PatS (Fig. 4 and Table 1). For similar reasons, secondary HetR proteins, though clearly deviant

from canonical HetR, probably retain function as transcriptional regulatory proteins, perhaps supplementing primary HetR or perhaps serving some other purpose. It should be noted that all of the organisms bearing secondary HetR proteins, except *Prochlorothrix hollandica* PCC 9006, are closely related to one another. That and the clustering of secondary HetR proteins (Supporting Fig. S2) is consistent with one or perhaps two acquisitions of the alternative form.

The exceptional filamentous strains (Fig. 6) include two strains of *Chlorogloeopsis* that lack PatX and three strains of *Anabaena* that have versions of PatX carrying RGTGR instead of RGSGR. These five strains are amongst the six cyanobacteria that carry HetN (if the HetN-like proteins of the *Chlorogloeopsis* strains are functional). Perhaps HetN partially substitutes for the function of PatX in these strains. It is also possible that *Chlorogloeopsis* strains need no HetN nor PatX at full strength to control HetR, since they are rarely in a filamentous state (Rippka *et al.*, 1979; Waterbury, 2006; Koch *et al.*, 2017). There is one more filamentous strain, *Crinalium epipsammum* PCC 9333, whose PatX protein has RGTGR in place of RGSGR and four closely related strains of Planktothrix with a motif of RGGGR. In the latter cases, the strains also possess a second gene similar in sequence characteristics and upstream sequence to conventional PatX (Supporting Fig. S8B) but in non-standard genetic contexts. In short, except in the case of *Crinalium epipsammum* PCC 9333, all filamentous cyanobacteria that lack PatX or have nonstandard PatX possess a protein that may conceivably compensate for the defect.

The regulatory regions associated with both *patS* and *patX* genes are consistent with their observed expression patterns and indicate that these characteristics may hold for cyanobacteria apart from the laboratory workhorse *Anabaena* PCC 7120. Both genes are preceded by conserved upstream regions with the following characteristics (Fig. 8 and Supporting Figs. S7 and S8). Both have DIF1 motifs (exact palindrome TCCGGA in the case of heterocyst-forming organisms, TCC[G/T]GA in the case of non-heterocyst-forming organisms) at a position that is probably -35 to the transcriptional start site, by analogy with the proven case of *Anabaena* PCC 7120 (Mitschke *et al.*, 2011). At the presumed -10 position there is another conserved motif: GTAGAGA (*patS*) or GTAnnAG (*patX*). In the case of *patX* in heterocyst-forming cyanobacteria, the DIF1 motif is preceded by an NtcA-binding site, which conceivably could mediate the observed nitrogen-dependence of high *patX* expression (Mitschke et al., 2011) or perhaps increase its level. Transcription from the *patX* promoter in *Anabaena* PCC 7120 responds to nitrogen deprivation in the same way (8 hrs vs 0 hrs) as transcription from the *patS* promoter, but the latter is an order of magnitude lower (Mitschke *et al.*, 2011). Similarly, the overall level of *patX* and *patS* RNA abundance follows the same pattern over 0, 6, 12, and 21 hours after deprivation, but the RNA abundance of *patS* is a few-fold higher than that of *patX* (Flaherty *et al.*, 2011). Perhaps the 3' end of *patX* mRNA is degraded more rapidly. Like other DIF1 transcriptional start sites, those associated with *patX* and *patS* are induced dependent on wild-type HetR (Mitschke *et al.*, 2011).

From these considerations, it is possible to propose a plausible sequence of evolutionary events that led to the cyanobacteria present today (Fig. 3). Multicellularity arose from a primordial unicellular state, and not long after the divergence of the *Pseudanabaena*, HetR and PatX entered the lineage, conferring some advantage to multicellular cyanobacteria (Schirrmeister, *et al.*,

2013). The juxtaposition of the two genes in most filamentous strains may reflect a primordial genetic linkage. The innovation of expensive but oxygen-resistant nitrogen fixation in heterocysts was enabled by tying the expression of HetR/PatX to nitrogen availability (through NtcA, directly or indirectly) on one hand and on the other hand to the regulation of heterocyst-related genes. The appearance of PatS made possible a greater degree of control, but the protein was lost in the common ancestor of the clade that includes *Anabaena cylindrica* PCC 7122. In the common ancestor of *Nostoc* PCC 7524, *Anabaena* PCC 7120, and its two closest relatives, an allele of a short-chain dehydrogenase/reductase appeared that had gained RGSGR within its sequence, resulting in a protein designated HetN. Subsequently, the RGSGR motif in PatX carried by the common ancestor of three of these strains mutated to RGTGR, leading possibly to diminished function of the protein. If one considers only the required elements of HetN (Corrales-Guerrero *et al.*, 2014) – its membrane associated N-terminus and its RGSGR motif – it resembles PatX and might partially substitute for it (but see an alternate view in Higa *et al.*, 2012).

Filamentarity is a polyphyletic trait, having been lost several times, every time with the concomitant loss of HetR and PatX or, in the cases of *Synechococcus* PCC 7002 and *Synechococcus* PCC 7335, the degradation (or repurposing) of HetR. This does not imply that HetR and PatX are required for filamentarity -- clearly not the case in *Anabaena* PCC 7120 (Buikema and Haselkorn, 1991 and I. Khudyakov, unpublished)—but indicates that this pair of proteins may regulate the multicellular behavior enabled by filamentarity. Fig. 3 interprets events as having a single acquisition of filamentarity, HetR, and PatX and multiple losses. It is more parsimonious to envision multiple acquisitions of filamentarity (and HetR and PatX), as others have suggested (Schirrmeister *et al.*, 2015), but if acquisition of filamentarity is more difficult than its loss, then simple parsimony may be a poor guide. The test is whether proteins such as HetR that are associated with filamentarity appear to have arisen in *Synechococcus* PCC 7002 and *Synechococcus* PCC 7335 by descent from a filamentous ancestor or by horizontal gene transfer. In the case of HetR, the phylogenetic tree (Supporting Fig. S2) is consistent with descent, but bootstrap support is weak. Stucken *et al.* (2010) listed 31 other proteins that were at the time associated specifically with filamentous strains. The much greater number of genome sequences available now reduces that number to only a few (J. Elhai, unpublished). , One of them, Alr4863, has orthologs in all filamentous strains except RicHH1 and orthologs in no unicellular strains, except *Synechococcus* PCC 7002, *Synechococcus* PCC 7335, and *Chamaesiphon minutus* PCC 6605 (another unicellular strain with a close filamentous relative). The phylogenetic tree of this protein is nearly superimposable upon the organismal tree (Fig. 3), after erasing the unicellular strains except *Synechococcus* PCC 7002, *Synechococcus* PCC 7335, and *Chamaesiphon minutus* PCC 6605 (result not shown), compatible with the lineal descent of filamentarity.

Our view of the regulation of heterocyst differentiation has been colored by the peculiarities of a single, atypical strain, *Anabaena* PCC 7120. While that strain possesses regulatory elements (HetR, PatS, and HetN) that fit well with a Turing/Meinhardt model of patterned differentiation, other heterocyst-forming cyanobacteria possess different elements (HetR, PatX, and maybe PatS), though the patterns of heterocysts in many of these are indistinguishable from those of

Anabaena PCC 7120. By the same token, it may be a mistake to view the role of the apparently primordial HetR and PatX, elements of a Turing/Meinhardt machine through the lens of spaced heterocyst differentiation. Its original function evidently evolved in filamentous strains without heterocysts, presumably related to a different sort of pattern advantageous to cyanobacteria with long filaments. One possibility is the formation of spaced necridia to promote hormogonia formation (Lamont, 1969; Nürnberg *et al.*, 2014).

Experimental Procedures

Strains, growth conditions and microscopy

Anabaena sp. strain PCC 7120 was grown in nitrate-replete BG-11 or combined nitrogen-free BG-11_0 medium (Rippka et al, 1979) at 30°C in constant fluorescent light. Its derivatives carrying replicative plasmids pRIAM810 or pRIAM971 were grown the presence of neomycin, 25 µg/ml for solid medium and 15 µg/ml for liquid medium. For heterocyst induction on agar plates, filaments from fresh streaks on BG-11 plates were patched on BG-11_0 plates. To induce heterocyst formation and/or follow *gfp* reporter induction in liquid medium the filaments from fresh streaks on BG-11 plates were transferred with sterile toothpicks into 96-well microtiter plate containing liquid BG-11_0 medium. Cells were examined by bright-field and fluorescence microscopy with a Zeiss Axio Imager A1 microscope equipped with HBO 100 mercury lamp source and Zeiss filter set 38HE (excitation: BP 470/40 nm, emission: BP 525/50 nm) for GFP fluorescence and filter set 14 (excitation: BP 510-560 nm, emission: LP 590 nm) for phycobilisome-induced autofluorescence. Images were recorded with a digital camera AxioCam HRc.

Cyanobacterial genomes and proteins

127 core cyanobacterial genomes, including plasmids, from all major groupings, were accessed along with their proteins through CyanoBIKE, an instance of BioBIKE (Elhai *et al.*, 2009; http://biobike.csbc.vcu.edu/) that focuses on cyanobacterial genomes. Additional cyanobacterial genomes were obtained from the NCBI. The origins and other characteristics of these genomes are shown in Supporting Table S1. Orthologous proteins from within BioBIKE were found as described below. Some additional HetR proteins were identified by querying the nonredundant protein database of NCBI through BlastP (Altschul *et al.*, 1997), with HetR from *Anabaena* PCC 7120 (Alr2339) as the query. Proteins found in this way were added to the set of HetR proteins if they matched at least 90% of the query and came either from a unicellular strain or from a strain with two or more proteins fitting the criteria.

Phylogenetic trees

Organismal trees were built by analyzing concatenated alignments of 29 proteins found in all 127 core cyanobacteria considered in this study. The names and coordinates of the orthologous proteins for each organism are given in Supporting Table S2. Most of the proteins were readily obtained by BioBIKE's ORTHOLOG-OF function, but in 16 cases (0.4% of the total number), a protein was not found in an organism, either because it had not been annotated or the ORF was broken by an apparent frame shift, either in Nature or in the sequencing and assembly of the genome. In such cases (noted in Supporting Table S2), the ORF was detected using TBlastN (Altschul *et al.*, 1997) via the SEQUENCE-SIMILAR-TO function (protein vs translated DNA) and repaired digitally. In addition to these, 21 proteins (0.6% of the total number) had start

codons apparently miscalled, truncating the annotated protein relative to other orthologs. In these cases the gene sequence was extended upstream to the presumably correct start codon matching those used by orthologs.

Alignments of each set of proteins (sets provided in Supporting Table S3) were made through Clustal W (Thompson *et al.*, 1994) accessed within BioBIKE and concatenated using an ad hoc BioBIKE script. The most informative columns were extracted using Gblocks 0.91b (Talavera *et al.*, 2007; http://molevol.cmima.csic.es/castresana/ Gblocks_server.html), and the final tree was made by PhyML 3.0 (Guindon *et al.*, 2010; http://www.atgc-montpellier.fr/phyml/), with LG as the substitution model, NNI as the type of tree improvement, and 100 bootstraps. The tree was visualized and manipulated using FigTree 1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/). Individual protein trees were made in an analogous fashion.

Identification of members of protein families

Orthologous proteins were obtained through BioBIKE's ORTHOLOG-OF function, which defines an ortholog by bidirectional best hit with a threshold of 10^{-10}. In other words, a protein A in organism X is defined as orthologous to protein B in organism Y if B is the best Blast hit of A against Y and A is the best Blast hit of B against X (E values < 10^{-10}). Amino acid sequences of HetR-like, HetN-like, NtcA-like sequences, and proteins encoded by genes typically flanking *patS* and *patX* genes were found within BioBIKE by ORTHOLOG-OF and SEQUENCE-SIMILAR (both protein vs protein and protein vs translated DNA). Identification was confirmed by examination of protein alignments. Exemplar proteins used to find the orthologs are shown in Supporting Tables S4, S5, and S6.

RGSGR-containing proteins were identified by a combination of several methods. First, a BioBIKE expression returned all ORFs (whether annotated or not) containing RG[S/G/T]GR. Candidates were excluded if they were encoded by ORFs that are out of frame within conserved genes.

Second, candidate PatS and PatX ORFs were sought without regard to genetic context by considering an exhaustive set of amino acid fragments from a genome in light of the the positional amino acid characteristics of a training set based on reliable PatS or PatX proteins. The exhaustive set was comprised of every possible amino acid fragment of a given length taken from a given genome translated in each of six possible reading frames. The positional characteristics of the relevant protein were captured by a position-specific scoring matrix (PSSM), a lookup table giving the adjusted probability of a given amino acid at a given position in a set of aligned sequences. The probabilities were adjusted in two ways. First, in order to compare each positional probability to chance, each probability in the PSSM was divided by the background frequency of the appropriate amino acid, given by BioBIKE's BACKGROUND-FREQUENCIES-OF function acting over all proteins in an organism. Prior to that, a constant number of counts (called pseudocounts) were added to all amino acid counts at all positions, to minimize the effect of low counts for a nucleotide owing to a small sample size. The number chosen was the square root of N divided equally amongst the 20 amino acids, where N is the number of sequences in the training set. Although this value is not uncommon (Higgs and Attwood, 2013), its only justification is that it is a reasonable compromise, valuing high frequency amino acids (e.g. RGSGR) while allowing for arbitrary variations. With this pseudocount value, a deviation from a unanimous position imposed a scoring penalty of -1.1 to -2.3 (lower for rare substituting amino acids like cysteine, higher for common amino acids like

leucine). The PSSM was calculated from the training set and target genome and applied to a translated genome sequence using the APPLY-PSSM-TO function

The function calculates a score for each of the several million ORF fragments extracted from the translated genome. The score, called the log odds, is the Log_{10} of the ratio:

$$(\text{Probability(region, given frequencies}_{\text{training set}}) / \text{Probability(region, given frequencies}_{\text{random}})$$

The number increases in magnitude as the region approaches the positional characteristics of the training set. A number is close to 0 is expected if the sequence of a segment arose by chance.

A third strategy was to refine the list produced by PSSM analysis, using the SEQUENCE-SIMILAR-TO with the MISMATCHES option to identify candidate ORFs with zero or one mismatch to RSGSR. In parallel, CONTEXT-OF and DESCRIPTION-OF genetic context and annotation of genes, making it possible to remove candidate ORFs within conserved genes.

Analysis of protein characteristics

Transmembrane domains were predicted with TMHMM 2.0 (TransMembrane-Hidden Markov Model) (Krogh and Sonnhammer, 2000; http://www.cbs.dtu.dk/services/TMHMM/), a program that compares the sequence characteristics of a given protein to those of a training set of 160 well studied membrane spanning regions of proteins from eukaryotes and prokaryotes. The program employs hidden Markov models, which use the propensities of regions of the training set to predict the likelihood that an amino acid follows a given amino acid string, much like using the high incidence "t" after "ich" to predict that one is probably looking at German text rather than English. The program reports membrane-spanning regions, but it is easily fooled by signal sequences.

Putative signal sequences (indicated in Fig. 6) were identified using SignalP 4.1 (Petersen *et al.*, 2011; http://www.cbs.dtu.dk/services/SignalP/). The program uses neural networks built from one of three data sets to distinguish signal sequences from non-signal sequences. Neural networks build decision-making processes from known exemplars. SignalP employs data sets using known signal sequences from human proteins (representing eukaryotes), from *Bacillus subtilis* (representing Gram-positive eubacteria), and from *Escherichia coli* (representing Gram-negative eubacteria). The networks may be trained positively with proven signal peptides and negatively with transmembrane sequences or trained only positively, relying on the user to assert that the input sequences have no confounding transmembrane sequences. The program also offers a strict threshold (default) and a permissive threshold (sensitive), or a knowledgeable user can specify any threshold. SignalP therefore offers 12 choices: 3 data sets x 2 training regimes x 2 levels of set sensitivity.

Cyanobacteria have cell wall characteristics similar to both Gram-negative and Gram-positive bacteria (Holczyk and Hansel, 2000), and they are phylogenetically no closer to *E. coli* than to *B. subtilis* (Hug, et al., 2016), so it is not obvious which neural network to choose. To address this question, we took the sequences from 162 proteins from *Synechocystis* PCC 6803 whose true N-termini had been determined (Sazuka *et al.*, 1999), 22 with apparent signal peptides and 140 whose N-terminus is the methionine at the predicted translational start site or the amino acid next to it. Running these proteins through the three neural networks of SignalP allowed us to assess the false positive and false negative rates and in this way judge the best of the 12 choices for determining cyanobacterial signal peptides. The Gram-positive data set (and, surprisingly, the

eukaryotic data set) outperformed the Gram-negative data set. The best condition was to use the Gram-positive data set with no training on transmembrane regions and high sensitivity (5% false negative and 1% false positive). We used this condition in predicting signal sequences but also show results from the same data set trained on transmembrane regions.

Analysis of upstream DNA sequence motifs

Sequences upstream from candidate *patX* genes were collected with BioBIKE's SEQUENCES-UPSTREAM-OF function. To search for DIF1 motifs upstream from candidate *patS* genes, 1000 nucleotides upstream from the start site were scanned for sequences with no more than one mismatch in TCCGGA.

To distinguish biologically functional matches from spurious matches, the characteristics of the DNA sequences surrounding the TCCGGA sites were compared to characteristics of a training set, 54-nucleotide sequences containing TCCGGA sites preceding *Anabaena* PCC 7120 genes known to be dependent on HetR (Mitschke *et al.*, 2011). The training set was used to construct a PSSM. In calculating the probabilities, only those positions in the aligned training were considered where the information in the training set exceeds 0.2 bits (see Fig. 8 for examples of information content). Other PSSMs were constructed in an analogous fashion. Some PSSMs were constructed based on training sets of variable length, owing to a gap of 17 or 18 nucleotides separating conserved regions. In such cases, sequences were forced to the same length by deleting when appropriate a nucleotide at the center of the gap region (after determining that this position has negligible information content). The efficacy of the procedure was supported by an analysis that compared scores DIF1 regions obtained as described above to a distribution of scores of the regions surrounding all genomic TCCGGA sites (almost all presumably random occurrences). See Supporting Fig, S9 for details.

The information (a measure of order) at a certain position in an alignment is defined as the difference between maximal disorder ($\mathbf{E_{max}}$) and disorder ($\mathbf{E}$) calculated at the given position, $\mathbf{E}$ is $-\Sigma\ \mathbf{p_i} \log_2 \mathbf{p_i}$ summed over all four nucleotides, $\mathbf{p_i}$ is the frequency of a given nucleotide at the position, $\log_2 \mathbf{p_i}$ is taken to be 0 when $\mathbf{p_i}$ is 0, and $\mathbf{E_{max}}$ is the maximum possible value of $\mathbf{E}$ and occurs when there is an even distribution of nucleotides ($\mathbf{p_i} = 1/4$), so $\mathbf{E_{max}}$ is (summed over the four nucleotides) $-\Sigma\ (1/4) \log_2 (1/4) = 2$. The maximum information therefore is 2, occurring when $\mathbf{p_i}$ for one nucleotide is 1 and for the other three, 0. Information for alignments was calculated using the INFORMATION-OF function of BioBIKE.

The information of positions within an alignment was visualized using WebLogo (Crooks *et al.*, 2004; http://weblogo.berkeley.edu/logo.cgi). Since the number of amino acids differs from the number of nucleotides, the maximum information value for amino acid sequences differs as well. That value is $-\Sigma\ (1/20) \log_2 (1/20) = 4.3$. For both nucleotide and amino acid logos, perfectly aligned positions may not have the maximum information value, because the program applies a correction for small sample size when the number of nucleotide sequences is less than 20 or amino acid sequences is less than 40.

Plasmid constructions and conjugations

Plasmid pRIAM810 is a mobilizable shuttle vector capable of replication in both *E. coli* and *Anabaena* and containing P_{petE}-*patX* for ectopic overexpression of *patX* gene. The plasmid was produced through several steps illustrated in Supporting Fig. S10. First, a plasmid, anp03226 (Kaneko et al., 2001), was obtained from C.P. Wolk that bears *patX* (*asl2332*) on *Anabaena* sp.

PCC 7120 chromosomal DNA from bp 2805907 to 2813409 cloned in the BamHI site of pUC18 with the gene antiparallel to the *lac* promoter. The plasmid was digested with EcoRI and self-ligated, excising most of the insert and producing pRIAM802, which contains *all2333* and *asl2332* on a 3047-nt insert. In parallel, the copper-regulated *petE* promoter (P_{petE}) was excised on a BamHI-EcoRI fragment from pPetE1 (Buikema and Haselkorn, 2001) and inserted between the BamHI and EcoRI sites of pBS SK. P_{petE} from the resulting plasmid was linked to a Sp^r/Sm^r cassette (Ω) by inserting the cassette on a 2 kb SmaI fragment taken from pAM684 (Ramaswamy et al., 1997) into the ScaI site at the 5′ end of the P_{petE} promoter, producing pRIAM806. The orientation of the *aad* gene in the omega cassette is antiparallel to P_{petE}. A fragment carrying Ω and P_{petE} was excised from pRIAM806 with BamH+EcoRI and moved into pRIAM802 cut with the compatible enzymes BglII (found naturally 231 nt upstream from *all2333*) and MfeI (304 nt from the 3' end of the gene, producing pRIAM808. In this way, the Ω-P_{petE} fragment replaced most of the 5′ portion of *all2333*, leaving *patX* under the control of its native promoter but placing the *petE* promoter further upstream. A SalI-EcoRI fragment from pRIAM808 with Ω-P_{petE}-*patX* was moved into pAM504, a cloning vector capable of replication in *Anabaena* (Wei et al., 1994), producing kanamycin-, neomycin-, streptomycin-, and spectinomycin-resistant pRIAM810.

Plasmid pRIAM971 is a mobilizable shuttle vector containing *patX* controlled by its native promoter and fused transcriptionally to GFP (P_{patX}-*patX'-gfp*). A fragment containing the 3′ end of *all2333* and the 5′ end (the first 21 codons) of *asl2332* (*patX*) was amplified by PCR from anp03226 with the primers 2333-F1 (5′-CAGCTTG*GTCG**A**C*GTTACGG with an engineered SalI site, one mismatch indicated in bold) and 2332-R1 (5′-GCTATT*CCC**G**GG*TAATCAGAAAC with an engineered SmaI site, one mismatch indicated in bold) and cloned into pAL-TA vector (Evrogen), producing pRIAM960. The insert from pRIAM960 was moved as a SalI-SmaI fragment into the corresponding sites of pAM1956 (Yoon & Golden, 1998), placing it upstream from a promoterless *gfp* to create kanamycin- and neomycin-resistant pRIAM971.

The shuttle plasmids were transformed into *E. coli* strain AM1359 (Yoon & Golden, 1998) containing a broad host range plasmid and a plasmid to provide methylation and mobilization functions and then conjugated into *Anabaena* PCC 7120 using standard protocols (Elhai et al., 1997).

Acknowledgments: We thank Laura Antonaru, Don Bryant, Jim Golden, Dennis Nürnberg, Doug Risser, and Karina Stucken for useful discussions. We also thank Peter Wolk for recognizing our common research directions and putting us together. The authors have no conflicts of interest to declare.

Author Contributions: JE and IK both conceived of the project and acquired, analyzed, and interpreted sequence data, first independently and then together. JE and IK each wrote major portions of the article.

831 REFERENCES

832 Altschul,S.F., Madden,T.L., Schaffer,A.A.,Zhang, J., Zhang, Z., Miller,W., and Lipman, D.J.
833 (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.
834 *Nucleic Acids Res* **25**: 3389-3402.

835 Antonaru, L.A., and Nürnberg, D.J. (2017) Role of PatS and cell type on the heterocyst spacing
836 pattern in a filamentous branching cyanobacterium. *FEMS Microbiol Lett* **364**: fnx154**.**

837 Bauer, C.C., Ramaswamy, K.S., Endley, S.,Scappino, L.A., Golden, J.W., and Haselkorn, R.
838 (1997) Suppression of heterocyst differentiation in *Anabaena* PCC 7120 by a cosmid carrying
839 wild-type genes encoding enzymes for fatty acid synthesis. *FEMS Microbiol Lett* **151**: 23-39.

840 Black, T.A., Cai, Y., and Wolk, C.P. (1993) Spatial expression and autoregulation of *hetR*, a
841 gene involved in the control of heterocyst development in *Anabaena*. *Mol Microbiol* **9**: 77-84.

842 Black, T.A., and Wolk, C.P. (1994) Analysis of a Het$^-$ mutation in *Anabaena* sp. strain
843 PCC 7120 implicates a secondary metabolite in the regulation of heterocyst spacing. *J Bacteriol*
844 **176**: 2282-2292.

845 Buikema, W.J., and Haselkorn, R. (1991a) Isolation and complementation of nitrogen fixation
846 mutants of the cyanobacterium *Anabaena* sp. strain PCC7120. *J Bacteriol* **173**: 1879-1885.

847 Buikema, W.J., and Haselkorn, R. (1991b) Characterization of a gene controlling heterocyst
848 differentiation in the cyanobacterium *Anabaena* 7120. *Genes Devel* **5**: 321-330.

849 Buikema, W.J., and Haselkorn, R. (2001) Expression of the *Anabaena hetR* gene from a copper-
850 regulated promoter leads to heterocyst differentiation under repressing conditions. *Proc Natl*
851 *Acad Sci USA* **98**: 2729-2734.

852 Cai, Y., and Wolk, C.P. (1997) *Anabaena* sp. strain PCC 7120 responds to nitrogen deprivation
853 with a cascade-like sequence of transcriptional activations. *J Bacteriol* **179**: 267-271.

854 Callahan, S.M., and Buikema, W.J. (2001) The role of HetN in maintenance of the heterocyst
855 pattern in *Anabaena* sp. PCC 7120. *Mol Microbiol* **40**: 941-950.

856 Chao, K.-M., and Zhang, L. (2009) Sequence comparison: theory and methods. In *Scoring*
857 *Matrices*, Chapter 8. Springer-Verlag, London. pp. 149-172.

858 Corrales-Guerrero, L., Mariscal, V., Flores, E., and Herrero, A. (2013) Functional dissection and
859 evidence for intercellular transfer of the heterocyst-differentiation PatS morphogen. *Mol*
860 *Microbiol* **88**: 1093-1105.

861 Corrales-Guerrero, L., Mariscal, V., Nürnberg, D.J., Elhai, J., Mullineaux, C.W., Flores, E., and
862 Herrero, A. (2014) Subcellular localization and clues for the function of the HetN factor
863 influencing heterocyst distribution in *Anabaena* sp. strain PCC 7120. *J Bacteriol* **196**: 3452-
864 3460.

865 Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004) WebLogo: a sequence logo
866 generator, *Genome Res*, 14: 1188-1190.

Dawson, J.P., Weinger, J.S., and Engleman, D.M. (2002) Motifs of serine and threonine can drive association of transmembrane helices. *J Mol Biol* **316**: 799-805.

Dong, Y., Huang, X., Wu, X.-Y., and Zhao, J. (2000) Identification of the active site of HetR protease and its requirement for heterocyst differentiation in the cyanobacterium *Anabaena* sp. strain PCC 7120. *J Bacteriol* **182**: 1575-1579.

Elhai, J., Vepritskiy, A., Muro-Pastor, A.M., Flores, E. and Wolk, C.P. (1997) Reduction of conjugal transfer efficiency by three restriction activities of *Anabaena* sp. strain PCC 7120. *J Bacteriol* **179:** 1998–2005.

Elhai, J., Taton, A., Massar, J.P., Myers, J.K., Travers, M., Casey, J., *et al.* (2009) BioBIKE: A web-based, programmable, integrated biological knowledge base. *Nucleic Acids Res* **37**: W28-W32.

Feldmann, E.A., Ni, S., Sahu, I.D., Mishler,C.H., Risser, D.D., Murakami, J.L., *et al.* (2011) Evidence for direct binding between HetR from *Anabaena* sp. PCC 7120 and PatS-5. *Biochemistry* **50**: 9212-9224.

Feldmann, E.A., Ni, S., Sahu, I.D., Mishler, C.H., Levengood, J.D., Kushnir, Y., *et al.*(2012) Differential binding between PatS C-terminal peptide fragments and HetR from *Anabaena* sp. PCC 7120. *Biochemistry* **51:** 2436-2442.

Flaherty, B.L., Van Nieuwerburgh, F., Head, S.R., and Golden, J.W. (2011) Directional RNA deep sequencing sheds new light on the transcriptional response of *Anabaena* sp. strain PCC 7120 to combined-nitrogen deprivation. *BMC Genomics* **12**: 332.

Flaherty, B.L., Johnson, D.B.F., and Golden, J.W. (2014) Deep sequencing of HetR-bound DNA reveals novel HetR targets in *Anabaena* sp. strain PCC7120. *BMC Microbiol* **13**: 255.

Flores, E., and Herrero, A. (1994). Assimilatory nitrogen metabolism and its regulation, In *The Molecular Biology of Cyanobacteria*. Bryant, D.A. (ed.). Kluwer Academic Publishers, Boston, Mass. p. 487-517.

Gerdtzen, Z.P., Salgado, J.C., Osses, A., Asenjo, J.A., Rapaport, I., and Andrews, B.A. (2009) Modeling heterocyst pattern formation in cyanobacteria. *BMC Bioinform* **10 (Suppl 6)**: 516.

Gierer, A., and Meinhardt, H. (1972) A theory of biological pattern formation. *Kybernetik* **12**: 30-39.

Gugger, M.F., and Hoffmann, L. (2004) Polyphyly of true branching cyanobacteria (Stigonematales). *Intl J Syst Evol Microbiol* **54**: 349-357.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**: 307-321,

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2017) PhyML 3.0: new algorithms, methods and utilities. [WWW document] URL http://www.atgc-montpellier.fr/phyml/

Henikoff, S., and Henikoff, J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* **89**: 10915-10919.

Herrero, A., Stavans, J., and Flores, E. (2016) The multicellular nature of filamentous heterocyst-forming cyanobacteria. *FEMS Microbiol Rev* **40**: 831-854.

Higa, K.C., Rajagopalen, R., Risser, D.D., Rivers, O.S., Tom, S.K., Videau, P., and Callahan, S.M. (2012) The RGSGR amino acid motif of the intercellular signalling protein, HetN, is required for patterning of heterocysts in *Anabaena* sp. strain PCC 7120. *Mol Microbiol* **83**: 682-693.

Higgs PG, and Attwood TK (2005). Bioinformatics and Molecular Evolution. Malden MA, Blackwell Publishing.

Holczyk, E., and Hansel, A. (2000) Cyanobacterial cell walls: news from an unusual prokaryotic envelope. *J Bacteriol* **182**: 1191-1199.

Howard-Azzeh, M., Shamseer, L., Schellhorn, H.E., and Gupta, R.S. (2014) Phylogenetic analysis and molecular signatures defining a monophyletic clade of heterocystous cyanobacteria and identifying its closest relatives. *Photosynth Res* **122**: 171-185.

Hu, H.-X., Jiang, Y.-L., Zhao, M.-X., Cai, K., Liu, S., Wen, B., *et al.* (2015) Structural insights into HetR-PatS interaction involved in cyanobacterial pattern formation. *Sci Reports* **5**: 16470.

Huang X., Dong Y., and Zhao J. (2004) HetR homodimer is a DNA-binding protein required for heterocyst differentiation, and the DNA-binding activity is inhibited by PatS. *Proc Natl Acad Sci U S A* **101**: 4848-4853.

Hug, L.A., Baker, B.B., Anantharaman, K, Brown, C.T., Probst, A.J., Castelle, C.J. *et al.* (2016) A new view of the tree of life. *Nature Microbiol* **1**: 1-6.

Janson, S., Wouters, J., Bergman, B., and Carpenter, E.J. (1999). Host specificity in the *Richelia*-diatom symbiosis revealed by *hetR* gene sequence analysis. *Environ Microbiol* **1**: 431-438.

Kaneko, T., Nakamura, Y., Wolk, C.P., Kuritz, T., Sasamoto, S., Watanabe, A., *et al.* (2001) Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. strain PCC 7120. *DNA Res* **8:** 205–213; 227-53.

Khudyakov, I.Y., and Golden, J.W. (2004) Different functions of HetR, a master regulator of heterocyst differentiation in *Anabaena* sp. PCC7120, can be separated by mutation. *Proc Natl Acad Sci USA* **101**: 16040-16045.

Kim, Y., Joachimiak, G., Ye, Z., Binkowski, T.A., Zhang, R., Gornicki, P., *et al.* (2011) Structure of transcription factor HetR required for heterocyst differentiation in cyanobacteria. *Proc Natl Acad Sci USA* **108**: 10109-10114.

Kim, Y., Ye, Z., Joachimiak, G., Videau, P.,Young, J., Hurd, K., *et al.* (2013) Structures of complexes comprised of *Fischerella* transcription factor HetR with *Anabaena* DNA targets. *Proc Natl Acad Sci USA* **110**: E1716-E1723.

Koch, R., and Kupczok, A., Stucken, K., Ilhan, J., Hammerschmidt, K., Dagan, T. (2017) Plasticity first: molecular signatures of a complex morphological trait in filamentous cyanobacteria. *BMC Evol Biol* **17**: 209.

Krogh, A., Larsson, B., von Heijne G., and Sonnhammer, E.L.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J Mol Biol* **305**: 567-580.

Kumar, K., Mella-Herrera, R.A., and Golden, J.W. (2010) Cyanobacterial heterocysts. *Cold Spring Harb Perspect Biol* **2**: a000315.

Lamont, H.C. (1969) Sacrificial cell death and trichome breakage in an Oscillatorian blue-green alga: the role of murein. *Arch Mikrobiol* **69**: 237-259.

Li, B., Huang, X., and Zhao, J. (2002) Expression of *hetN* during heterocyst differentiation and its inhibition of *hetR* up-regulation in the cyanobacterium *Anabaena* sp. PCC 7120. *FEBS Lett* **517**: 87-91.

Li, X., Sandh, G., Nenninger, A., Muro-Pastor, A.M., and Stensjö, K. (2015) Differential transcriptional regulation of orthologous *dps* genes from two closely related heterocyst-forming cyanobacteria. *FEMS Microbiol Lett* **362**: fnv017.

Maldener, I., Summers, M.L., and Sukenik, A. (2014) Cellular differentiation in filamentous cyanobacteria. In *The Cell Biology of Cyanobacteria*. Flores, E., and Herrero, A. (eds). Caister Academic Press. pp. 263-291.

Marcon, L., and Sharpe, J. (2012) Turing patterns in development: What about the horse part? *Curr Opin Genet Dev* **22**: 578-584.

Meeks, J.C., Elhai, J., Thiel, T., Potts, M., Larimer, F., Lamerdin, J., *et al.*. (2002) An overview of the genome of *Nostoc punctiforme*, a multicellular, symbiotic cyanobacterium. *Photosyn Res* **70**: 85-106.

Meinhardt, H. (2008) Models of biological pattern formation: From elementary steps to the organization of embryonic axes. *Curr Top Dev Biol* **81**: 1-63.

Mitschke, J., Vioque, A., Haas, F., Hess, W.R., and Muro-Pastor, A.M. (2011) Dynamics of transcriptional start site selection during nitrogen stress-induced cell differentiation in *Anabaena* sp. PCC7120. *Proc Natl Acad Sci USA* **108**: 20130-20135.

Monera, O. D., Sereda, T.J., Zhou, N.E., Kay, C.M., and Hodgesis, T.S. (1995) Relationship of sidechain hydrophobicity and α-helical propensity on the stability of the single-stranded amphipathic α-helix. *J Peptide Sci* **1**: 319-329.

Muñoz-Garcia, J., and Ares, S. (2016) Formation and maintenance of nitrogen-fixing cell patterns in filamentous cyanobacteria. *Proc Natl Acad Sci USA* **113**: 6218-6223.

Muro-Pastor, A.M. (2014). The heterocyst-specific NsiR1 small RNA is an early marker of cell differentiation in cyanobacterial filaments. *mBio* **5**:e01079-14.

Muro-Pastor, A.M., Brenes-Álvarez, and M., Vioque, A. (2017) A combinatorial strategy of alternative promoter use during differentiation of a heterocystous cyanobacterium. *Environ Microbiol Rep* **9**: 449-458.

Muro-Pastor, A.M., Valladares, A., Flores, E., and Herrero, A. (2002) Mutual dependence of the expression of the cell differentiation regulatory protein HetR and the global nitrogen regulator NtcA during heterocyst development. *Mol Microbiol* **44**: 1377-1385.

National Center for Biotechnology Information (2017) Prokaryotic RefSeq genome re-annotation project. [WWW document] URL https://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/reannotation/

Nayar, A.S., Yamaura, H., Rajagopalan, R., Risser, D.D., and Callahan, S.M. (2007) FraG is necessary for filament integrity and heterocyst maturation in the cyanobacterium *Anabaena* sp. strain PCC 7120. *Microbiol* **153**: 601-607.

Nürnberg, D.J., Mariscal, V., Parker, J., Mastroianni, G., Flores, E., Mullineaux, C.W. (2014) Branching and intercellular communication in the Section V cyanobacterium *Mastigocladus laminosus*, a complex multicellular prokaryote. *Mol Microbiol* **91**: 935-949.

Petersen, T. N., Brunak, S., von Heijne, G., and Nielsen, H. (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods* **8**: 785-786.

Picossi, S., Flores, E., and Herrero, A. (2014) ChIP analysis unravels an exceptionally wide distribution of DNA binding sites for the NtcA transcription factor in a heterocyst-forming cyanobacterium. *BMC Genomics* **15**: 22.

Rajagopalan, R., and Callahan, S.M. (2010) Temporal and spatial regulation of the four transcription start sites of *hetR* from *Anabaena* sp. strain PCC 7120. *J Bacteriol* **192**: 1088-1096.

Ramaswamy, K.S., Carrasco, C.D., Fatma, T., and Golden, J.W. (1997) Cell-type specificity of the *Anabaena fdxN*-element rearrangement requires *xisH* and *xisI*. *Mol Microbiol* **23:** 1241–1250.

Rippka, R., Deruelles, J., Waterbury, J.B., Herdman, M., and Stanier, R.Y. (1979) Generic assignments, strain histories and properties of pure cultures of cyanobacteria. *J Gen Microbiol* **111**: 1-61.

Risser, D.D., and Callahan, S.M. (2007) Mutagenesis of *hetR* reveals amino acids necessary for HetR function in the heterocystous cyanobacterium *Anabaena* sp. strain PCC7120. *J Bacteriol* **189**: 2460-2467.

Risser, D.D., and Callahan, S.M. (2009) Genetic and cytological evidence that heterocyst patterning is regulated by inhibitor gradients that promote activator decay. *Proc Natl Acad Sci USA* **106**: 19884-19888.

Rivers, O.S., Videau, P., and Callahan, S.M. (2014) Mutation of *sepJ* reduces the intercellular signal range of a *hetN*-dependent paracrine signal, but not of a *patS*-dependent signal, in the filamentous cyanobacterium *Anabaena* sp. strain PCC7120. *Mol Microbiol* **94**: 1260-1271.

Sánchez-Baracaldo, P., Ridgwell, A., and Raven, J.A. (2014) A neoproterozoic transition in the marine nitrogen cycle. *Curr Biol* **24**: 6520657.

Saw, J.H.W., Schatz, M., Brown, M.V., Kunkel, D.D., Foster, J.S., Shick, H., *et al.* (2013) Cultivation and complete genome sequencing of *Gloeobacter kilaueensis* sp. nov., from a lava cave in Kīlauea Caldera, Hawai'i. *PloS ONE* **8**: e76376.

Sazuka, T., Yamaguchi, M., and Ohara, O. (1999) Cyano2Dbase updated: Linkage of 234 protein spots to corresponding genes through N-terminal microsequencing. *Electrophoresis* **20**: 2160-2171.

Schirrmeister, B.E., de Vos, J.M., Antonelli, A., and Bagheri, H.C. (2013) Evolution of multicellularity coincided with increased diversification of cyanobacteria and the Great Oxidation Event. *Proc Natl Acad Sci USA* **110**: 1791-1796.

Schirrmeister, B.E., Gugger, M., and Donoghue, P.C.J. (2015) Cyanobacteria and the great oxidation event: evidence from genes and fossils. *Paleontol* **58**: 769-785.

Shih, P.M., Wu, D., Latifi, A., Axen, S.D., Fewer, D.P., Talla, E., *et al.* (2014) Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proc Natl Acad Sci USA* **110**: 1053-1058.

Stucken, K., John, U., Cembella, A., Murillo, A.A., Soto-Liebe, K., Fuentes-Valdés, J.J., *et al.* (2010) The smallest known genomes of multicellular and toxic cyanobacteria: comparison, minimal gene sets for linked traits and the evolutionary implications. *PLoS One* 5:e9235.

Talavera, G., and Castresana, J. (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol* **56**: 564-577.

Thompson, J.D., Higgins, D.G., and Gibson, T. J. (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**: 4673-4680.

Toyoshima, M., Sasaki, N.V., Fujiwara, M., Ehira, S., Ohmori, M., and Sato, N. (2010). Early candidacy for differentiation into heterocysts in the filamentous cyanobacterium *Anabaena* sp. PCC 7120. *Arch Microbiol* **192**: 23-31.

Tumer, N.E., Robinson, S.J., and Haselkorn, R. (1983) Different promoters for the *Anabaena* glutamine synthetase gene during growth using molecular or fixed nitrogen. *Nature* **306**: 337-342.

Turing, A.M. (1952) The chemical basis of morphogenesis. *Phil Trans Royal Soc B* **237**: 37-72.

Uyeda, J.C., Harmon, L.J., and Blank, C.E. (2016) A comprehensive study of cyanobacterial morphological and ecological evolutionary dynamics through deep geologic time. *PloS ONE* **11**: e0162539.

Valladares, A., Flores, E., and Herrero, A. (2016) The heterocyst differentiation transcriptional regulator HetR of the filamentous cyanobacterium *Anabaena* forms tetramers and can be regulated by phosphorylation. *Mol Microbiol* **99**: 808-819.

Videau, P., Ni, S., Rivers, O.S., Ushijima,B., Feldmann, E.A., Cozy, L.M., *et al.* (2014a) Expanding the direct HetR regulon in *Anabaena* sp. strain PCC 7120. *J Bacteriol* **196**: 1113-1121.

Videau, P., Oshiro, R.T., Cozy, L.M., and Callahan, S.M. (2014b) Transcriptional dynamics of developmental genes assessed with an FMN-dependent fluorophore in mature heterocysts of *Anabaena* sp. strain PCC 7120. *Microbiol* **160**: 1874-1881.

Voβ, B., Bolhuis, H., Fewer, D.P., Kopf, M., Möke, F., Haas, F., *et al.* (2013) Insights into the physiology and ecology of the brackish-water-adapted cyanobacterium *Nodularia spumigena* CCY9414 based on a genome-transcriptome analysis. *PLoS ONE* **8:** e60224.

Wang, H., Sivonen, K., Rouhiainen, L., Fewer, D.P., Lyra, C., Rantala-Ylinen, A., *et al*. (2012) Genome-derived insights into the biology of the hepatotoxic bloom-forming cyanobacterium *Anabaena* sp. strain 90. *BMC Genomics* **13**: 613.

Waterbury, J.B. (2006) The cyanobacteria -- isolation, purification, and identification. In *The Prokaryotes: Bacteria: Firmicutes, Cyanobacteria*, 3rd edition, Vol. 4. Dworkin, M., Falkow, S., Rosenberg, E., Schleifer, K-H, Stackebrandt, E. (eds). Springer-Verlag, New York. pp. 1053-1073.

Wei, T.F., Ramasubramanian, T.S., and Golden J.W. (1994) *Anabaena* sp. strain PCC 7120 *ntcA* gene required for growth on nitrate and heterocyst development. *J Bacteriol* **176:** 4473-82.

Wu, X. Liu, D., Lee, M.H., and Golden, J.W. (2004) *patS* minigenes inhibit heterocyst development of *Anabaena* sp. strain PCC 7120. *J Bacteriol* **186**: 6422-6429.

Yoon, H.- S., and Golden, J. W. (1998) Heterocyst pattern formation controlled by a diffusible peptide. *Science* **282**: 935-938.

Yoon, H.-S., and Golden, J.W. (2001) PatS and products of nitrogen fixation control heterocyst pattern. *J Bacteriol* **183**: 2605-2613.

Yoon, H.-S., Lee, M.H., Xiong, J., and Golden, J.W. (2003) *Anabaena* sp. strain PCC 7120 *hetY* gene influences heterocyst development. *J Bacteriol* **185**: 6995-7000.

Zhang, J.-Y., Chen, W.-L., and Zhang, C.-C. (2009) *hetR* and *patS*, two genes necessary for heterocyst pattern formation, are widespread in filamentous nonheterocyst-forming cyanobacteria. *Microbiol* **155**: 1418-1426.

FIGURE LEGENDS

Fig. 1. Model and proposed realization of a reaction-diffusion system. (**A**) Generic Turing/Meinhardt model. A molecule **R** regulates some action of interest. It also exerts positive feedback on its own synthesis or activity and increases the synthesis or activity of a second molecule, **S**, a suppressor of **R**. **S** is able to diffuse, while **R** remains at the site of synthesis. (**B**) Model of regulation of heterocyst differentiation in *Anabaena* PCC 7120. HetR plays the role of the R morphogen, here activating heterocyst differentiation. PatS and HetN collectively play the role of the S, suppressing HetR activity at different times during development. NtcA monitors nitrogen status. (**C**) Outcome of the proposed model -- filaments of *Anabaena* PCC 7120 grown in the absence of fixed nitrogen. Green cells are vegetative cells, capable of photosynthesis. Cells stained with Alcian Blue are heterocysts, incapable of photosynthesis but sites of nitrogen fixation. Photo courtesy of AV Matveyev.

Fig. 2. Overview of phylogenetic tree, rooted by *Gloeobacter violaceus* PCC 7421. See *Experimental procedures* section for details on construction and Fig. 3 for an expanded version of the tree. Strains highlighted in red are unicellular, blue heterocyst-forming, green other filamentous strains, and pink picocyanobacteria. The tree can be interpreted in multiple ways with regards to switches between unicellularity and filamentarity. What is shown is not the interpretation with the fewest switches but one that proceeds from the hypothesis that filamentarity arose only once. Group numbers correspond to those used by Howard-Azzeh *et al.* (2014). Arrows indicate proposed acquisition of proteins, and stars indicate proposed loss of proteins.

Fig. 3. Phylogenetic tree of cyanobacterial genomes and presence of key regulatory proteins. See Fig. 2 for phylogenetic context and the *Experimental procedures* section for a description of how the tree came about. Colored circles on the right indicate the presence in the genome of A=NtcA, R=HetR, N=HetN, S=PatS, X=PatX. Small circles indicate protein assignments for which there is doubt owing to synteny concerns, and triangles indicate proteins with differences with respect to RGSGR of HetN or PatX or two of the conserved residues of HetR (see text). Asterisks indicate genomes that have a surprising presence of HetR or a surprising absence of PatX. Yellow circles and orange circles in the tree indicates bootstrap support of at least 90% and 67%, respectively. Group numbers correspond to those used by Howard-Azzeh *et al.* (2014). (**A**) Heterocyst-forming cyanobacteria. One strain (*Raphiodopsis brookii* D9) does not make functional heterocysts, but shows evidence by electron microscopy of differentiation of terminal cells (Stucken *et al.*, 2010). (**B**) Non-heterocyst-forming filamentous cyanobacteria and unicellular cyanobacteria.

Fig. 4. Alignment of HetR proteins. 91 HetR protein sequences were aligned, including 40 filamentous, heterocyst-forming strains (highlighted blue), 43 filamentous, non-heterocyst-forming strains (highlighted green, except for 9 secondary copies as defined in the text, which are highlighted gray), and 7 unicellular strains (highlighted red). The order of the organisms is the same as in the phylogenetic tree of HetR (Supporting Fig. S2), and their full names are given in Fig. 3. Columns in which no more than three mutational events are evident from the primary filamentous HetR sequences are colored in green when the amino acid agrees with the consensus, cyan when it is a conservative substitution as defined by a BLOSUM80 (Henikoff

and Henikoff, 1992; Chao and Zhang, 2009) score of 2 or greater, gray when the substitution is not conservative but nonetheless represents a substitution of one of the six most hydrophobic residues (Monera *et al.*, **1995**) with another, or otherwise pink. The three lines at the top of the alignment (and repeated at the bottom) indicate residues for which there is evidence concerning functional importance. The top line (H) indicates whether at least one mutant residue affects heterocyst differentiation (red if the frequency of heterocysts markedly decreases, blue if it markedly increases, green if it does not change). The second line (A) indicates whether at least one mutant residue affects an in vitro assay for DNA binding (red) or PatS binding (blue). The residue is green if the assay of the mutant HetR gives a similar result as wild-type HetR. The bottom line (S) indicates whether an analysis of the structure of a crystalized HetR protein indicates binding of the residue to DNA (red), to PatS (blue), or to another HetR subunit (gray) The full alignment is given in Supporting Fig. S3, along with evidence for the functional assertions.

Fig. 5. Amino acid sequences of candidate PatS proteins**.** Amino acid sequences suspected of encoding a functional PatS protein were identified as described in the text, using sequence and contextual cues. The RGSGR motif is highlighted in dark green and a conserved preceding glutamate residue in light green. Gray and blue highlighted letters indicate hydrophobic and very hydrophobic amino acid side chains at pH7, respectively (Monera *et al.*, **1995**). Red letters indicate alternative start sites, with possible N-terminal extensions indicated in gray font. Despite the seemingly straightforward experiment of Corrales-Guerrero et al (2013), there is considerable doubt regarding the most active start codon for *patS* of *Anabaena* PCC 7120 (Ana7120), never mind the other instances of *patS*. See Supporting Table S7, which provides evidence that translation starts primarily at the valine codon (producing an 11-amino acid PatS) and to a lesser extent at the second methionine codon (producing a 13-amino acid PatS). The lower case italicized sequence of *Fischerella thermalis* PCC 7521 (Fis7521) is the virtual translation that removes a one-nucleotide insertion relative to the sequences of other *Fischerella*. Arrows indicate the position and direction of flanking genes (not to scale): dihydroorotase (blue), patatin (green), and HetY (red). Sequences lacking contextual support are marked with asterisks. *Mastigocoleus testarum* BC008 (Mas008) has two identical sequences. Organismal abbreviations are explained in Fig. 3.

Fig. 6. Amino acid sequences of candidate PatX proteins. Amino acid sequences suspected of encoding a functional PatX protein were identified as described in the text, using sequence and contextual cues. Amino acids are colored as described in Fig. 5. In addition, proline residues are highlighted in yellow. Arrows indicate the position and direction of flanking genes (not to scale): *hetR* (blue), *sepJ* (cyan), FAD-dependent oxidoreductase (DH; green), conserved hypothetical (Hyp; light pink), methyltransferase (MTase; dark pink), and *glnA* (red). Sequences lacking contextual support are marked with asterisks. Nicknames of the organisms are followed by a symbol indicating the presence of a N-terminal signal peptide sequence as predicted by SignalP (see *Experimental procedures*): # (present, strict conditions), + (present, but only if possible transmembrane regions are ignored), o (absent), and ~ (within 10% of threshold). Organismal abbreviations are explained in Fig. 3. (**A**) Heterocyst-forming cyanobacteria. (**B**) Non-heterocyst-forming filamentous cyanobacteria and unicellular cyanobacteria.

Fig. 7. Conserved amino acid residues in PatS and PatX. Sequence logos representing the amount of information (associated with the degree of conservation) are shown for sequences of (**A**) PatS, (**B**) PatX (heterocyst-forming cyanobacteria), and (**C**) PatX (other cyanobacteria). Residues are colored blue (positively charged), red (negatively charged, green (polar), and black (non-polar). Variable spacing between clusters of aligned amino acids are shown. See Figs. 6 and 7 for amino acid sequences of each protein.

Fig. 8. Conserved nucleotides upstream from PatS, PatX, and DIF1-motif-containing regions. Sequence logos representing the amount of information (associated with the degree of conservation) are shown for sequences upstream from (**A**) *patS*, (**B**) *patX* (heterocyst-forming cyanobacteria), (**C**) *patX* (other cyanobacteria), and (**D**) HetR- and N-regulated genes of *Anabaena* PCC 7120. Variable spacing between clusters of aligned nucleotides are shown as are DIF1 and NtcA-binding motifs. The approximate position of transcriptional initiation for the appropriate gene from *Anabaena* PCC 7120 (Mitschke *et al.*, 2011) is shown by an arrow. See Supporting Figs. S5 – S7 for nucleotide sequences of each upstream sequence.

Fig. 9. Kinetics of P_{patX}-*gfp* and P_{patS}-*gfp* reporters expression in *Anabaena* PCC 7120.
Wild-type *Anabaena* PCC 7120 carrying pRIAM970 with the reporter fusion P_{patX}-*gfp* or pAM830 (Yoon and Golden, 1998) with the reporter fusion P_{patS}-*gfp* was grown on neomycin-containing BG-11 plate and then transferred to combined nitrogen-free liquid BG-11_0 medium. Micrographs were taken 6 h and 16 h after nitrogen step down. Images correspond to phycobilisome-induced red autofluorescence (left panel) and GFP fluorescence (right panel). Note the reduced autofluorescence in developing heterocysts at 16 h. Arrowheads point to cells with high GFP fluorescence, presumably (at 6 h) potential (but not committed) proheterocysts and (at 16 h) developing proheterocysts (note the reduced autofluorescence in some of the indicated cells).

Fig. 10. Ectopic overexpression of *patX* on diazotrophic growth and heterocyst differentiation in *Anabaena* PCC 7120.
Wild-type *Anabaena* PCC 7120 carrying a control plasmid pAM1956 (A, C) or P_{petE}-*patX*-containing pRIAM810 (B, D) was grown on nitrate-containing BG-11 plate with 25 µg/ml of neomycin and then transferred in combined nitrogen-free liquid BG-11_0 (A, B) or on solid BG-11_0 (C, D) medium. Micrographs were taken 3 days (A, B) and 4 days (C, D) after nitrogen step down. Arrowheads indicate heterocysts.

Table 1: Frequency of mutations in conserved amino acids in HetR

Function[a]	Number of sites[b]	Primary HetR's (74)[c]		Secondary HetR's (9)[c]		Unicellular strains (2)[c]		Closest relative (2)[c]	
		Mutations[d]	Rate[e]	Mutations[d]	Rate[e]	Mutations[d]	Rate[e]	Mutations[d]	Rate[e]
DNA-binding	19 (11%)	3 (3 + 0) [14]**	0.002	13 (8 + 5) [28]*	0.076	1 (0 + 1) [8]*	0.026	0 (0 + 0) [3]	0.000
PatS-binding	7 (4%)	5 (5 + 0) [5]	0.010	4 (4 + 0) [10]	0.063	6 (3 + 3) [3]†	0.429	1 (1 + 0) [1]	0.071
Subunit interaction	14 (8%)	9 (6 + 3) [10]	0.009	29 (11 + 18) [20] †	0.230	10 (5 + 5) [6] †	0.357	3 (2 + 1) [2]	0.107
No information	132 (77%)	99 (55 + 44)	0.010	192 (70 + 122)	0.162	52 (25 + 27)	0.197	20 (13 + 7)	0.076
TOTAL	172 (100%)	116 (69 + 47)	0.009	239 (93 + 146)	0.154	69 (33 + 36)	0.201	24 (16 + 8)	0.070

[a] Amino acid residues associated with different functions. See Supporting Information Fig. 3 for evidence

[b] Conserved amino acid residues in each functional class. See Fig. 4 for definition of "conserved".

[c] Primary HetR's are those HetR that are in filamentous cyanobacteria and are phylogenetically most related, as described in the text. Secondary HetR's are very similar proteins in organisms with primary HetR's. There are two unicellular strains (*Synechococcus* PCC 7002 and *Synechococcus* PCC 7335) considered here bearing HetR-like proteins and two strains most closely related to them (*Leptolyngbya* PCC 7376 and *Leptolyngbya* PCC, respectively) .

[d] Number of mutational events in conserved sites of each functional class. The first number represents conserved amino acid changes and the second non-conserved, as described in Fig. 4. Events were estimated with guidance from the organismal (Fig. 3) and HetR (Suppl. Fig. 2) phylogenetic trees. The second and third numbers (in parentheses) indicate the number of conserved and non-conserved substitutions, respectively. The last number (in brackets) indicates the expected number of substitutions, if the total number of mutations for the class (e.g. primary HetR's) were distributed proportionally to the number of sites in each functional category relative to the number of mutations in the *No information* category. For example, the expected number of mutations in Primary HetR's at amino acid positions related to DNA-binding would be 99 (19/132). Values that are significantly less than expected according to a Chi-squared test are marked with one asterisk ($p < 0.05$) or two ($p < 0.01$). Values that are significantly more than expected are marked with a dagger ($p < 0.05$).

[e] The rate is defined as the number of mutational events divided by the total number of possible events (the number of sites in the functional category times the number of proteins in the class.