REGULATORY ELEMENTS OF MARINE CYANOBACTERIA

SZYMON M. KIEŁBASA[1] HANSPETER HERZEL[2]
kielbasa@molgen.mpg.de h.herzel@biologie.hu-berlin.de

ILKA M. AXMANN[2]
i.axmann@biologie.hu-berlin.de

[1] *MPI MG, Max Planck Institute for Molecular Genetics, Ihnestrasse 73, D-14195 Berlin, Germany*
[2] *ITB, Institute for Theoretical Biology, Humboldt University of Berlin, Invalidenstrasse 43, D-10115 Berlin, Germany*

The free-living, oxyphototroph bacteria of the group of *Prochlorococcus* populate widely the oceans. Genomic information of nine marine cyanobacteria was used to predict signals essential for regulation. We implemented a pipeline that automatically calculates BLASTp alignments of query genomes, selects a representative subset of orthologs and predicts motifs conserved in their upstream sequences. Next, similar motifs are clustered into groups which could contain profiles recognized by different transcription factors. The phylogenetic footprinting pipeline revealed a minimal conserved set of putative transcription factors, binding sites and regulons for the chosen marine cyanobacterial genomes. DNA-binding motifs for NtcA and LexA were correctly identified. The relevance of transcriptional regulation of predicted *cis* elements was supported experimentally.

Keywords: phylogenetic footprinting; transcription factor binding sites; marine cyanobacteria.

1. Introduction

Photosynthetic bacteria such as *Prochlorococcus* and *Synechococcus* belong to the most important primary producers within the oceans. The genus *Prochlorococcus* is often present at high abundances with more than 10^5 cells per ml in nutrient-poor areas of the world's oceans splitting up in two major ecotypes – one being represented by the high-light-adapted (HL) strains such as Med4, the other by low-light-adapted (LL) strains SS120 and MIT9313 [24, 26, 36]. Nevertheless, on the basis of their ribosomal DNA similarity different ecotypes would be recognized as a single species as their rDNA sequences differ by less than 3% [15]. At the molecular level only small pieces are known about the regulatory network of marine cyanobacteria and genome-wide studies about co-regulated genes (regulons) controlled by *trans*-acting transcription factors (TFs) and their *cis* encoded DNA-binding sites do not exist. Only a few putative TF binding sites have been analysed: one for the CRP-like regulator NtcA (TGT-N_{10}-ACA) [25, 30] known to mediate nitrogen control in cyanobacteria, and a motif of putative phosphate regulator PhoB (TTAACCTT-N_3-TTAACCAT) [29]. The existance of a LexA site was suggested but not shown

by [22]. Knowledge about further *cis* elements on DNA is still rare.
To get insights into the core network of regulatory elements of multiple related species, phylogenetic footprinting is the major method. Thereby candidate regulatory elements are found by searching for conserved motifs upstream of orthologous genes from closely related species. Sequence similarity is the foundation for this computational method assuming that mutations within functional regions of genes accumulate slower than mutations in regions without sequence-specific function [35]. The phylogenetic footprinting algorithms can be divided in three parts: defining orthologous gene sequences for comparison; aligning the promoter sequences of orthologous genes; identifying segments of significant conservation. The great power of phylogenetic footprinting algorithms has been demonstrated for organisms of all kingdoms of life as the prediction of transcription regulatory sites in diverse bacterial families [23, 37], yeast [5], mouse and human [13, 20]. Reviews of methods and available resources are given by numerous articles [4, 6, 10, 31, 35]. Thereby, the initial and maybe the most difficult decision is choosing a set of genomes with the appropriate evolutionary distance of the sequences. The genomes of nine highly related but likely differentially adapted marine *Prochlorococcus* strains may represent the right distance to obtain meaningful predictions of *cis* elements.
Thus, we analyzed nine marine *Prochlorococcus* genomes and we predicted a conserved transcriptional regulatory network. For the first time, a minimal conserved core set of transcription factors, their binding sites and regulons can be suggested for the smallest known photosynthetic organism. DNA-binding motifs for NtcA and LexA were identified and several new regulatory motifs were predicted. A weak singnal corresponding to a third known motif ArsR has been observed. The importance of transcriptional regulation of two predicted *cis* elements NtcA and LexA was supported by experimental results of transcription initiation sites.

2. Materials and methods

2.1. *Computational part*

We performed a systematic intergenomic comparison to detect similar transcription factor binding sites conserved in upstream regions of orthologous genes. A pipeline implemented with BioMinerva framework [18] was used to integrate genome data and third party software tools in order to identify orthologous genes, align their promoter sequences and later to compare the alignments and interpret their similarity as a signal of regulation by a transcription factor.
Nine genomes of *Prochlorococcus* sp. were downloaded from NCBI GenBank server[a]. Tab. 1 summarizes the properties of retrieved sequences and their annotations. In order to build gene families we extracted the gene protein sequences from all studied species. Those sequences were next aligned against themselves using the BLASTp [1] algorithm run with the default parameters. The outcoming alignments can be

[a] `ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/`, version of February 2007

Table 1. Overview of size, annotated number of protein-coding genes (CDS), GC content, light optima (HL–high light; LL–low light) and numbers of annotated σ and transcription factors (TFs) of nine studied *Prochlorococcus* genomes.

name	accession	size	CDS	GC%	adaptation	σ	TFs
AS9601	CP000551	1669886	1921	31.32	HL	5	17
CCMP1375	AE017126	1751080	1882	36.44	LL	5	20
Med4	BX548174	1657990	1716	30.80	HL	5	23
MIT9313	BX548175	2410873	2273	50.74	LL	8	29
MIT9303	CP000554	2682675	2997	50.01	LL	8	30
MIT9312	CP000111	1709204	1809	31.21	HL	4	19
MIT9515	CP000552	1704176	1906	30.79	HL	5	19
NATL1A	CP000553	1864731	2193	34.98	LL	5	18
NATL2A	CP000095	1842899	1890	35.12	LL	5	19

understood as a graph representing evolutionary similarities between the studied genes. This graph was then processed by Markov Cluster Algorithm (MCL) [9] (an algorithm for unsupervised graph clustering based on simulation of stochastic flow in graphs) leading to a list of gene families.

We assume that genes belonging to one family are probably regulated in similar manner although they belong to different organisms. This assumption we interpret as a high chance to detect similar regulatory binding sites of the same transcription factor in corresponding promoter regions. Therefore, we predict transcription factor binding sites for each family separately in the following way. For all genes of a family we prepare their upstream DNA sequences (till the next gene or up to 300 nt of length). If an upstream region is shorter than 50 nt we assume that the gene belongs to a larger operon and we exclude it from further analysis. Afterwards, the set of upstream regions is processed by GLAM [11] – a method calculating the best possible gapless local alignment of multiple sequences with automatic determination of alignment width. If a good alignment is found it can represent a set of sites bound by a transcription factor. Since typical regulatory sites are not long we limit the maximum alignment length to 20.

As a result we obtain multiple alignments for each gene family. Next, each such alignment is converted into a position specific count matrix (PSCM) understood as a profile recognized by a potential regulating transcription factor. Since a typical transcription factor should regulate genes of more than one family we search for similar profiles calculated for different gene families. This step is carried out by a PSCMs comparison method (wmCompare [19] which bases on correlation between locations of binding sites predicted for a pair of PSCMs). The outcome of the method is an ordered list containing all matrices pairs and their similarities.

We take the most similar pairs above a chosen threshold of matrices similarity and estimate the biological significance of the choice. For this purpose we take the original matrices and shuffle their contents. In this process each position specific count matrix is converted into another one with randomly reordered positions (but still with the same size, quality, information content and GC-content). In general,

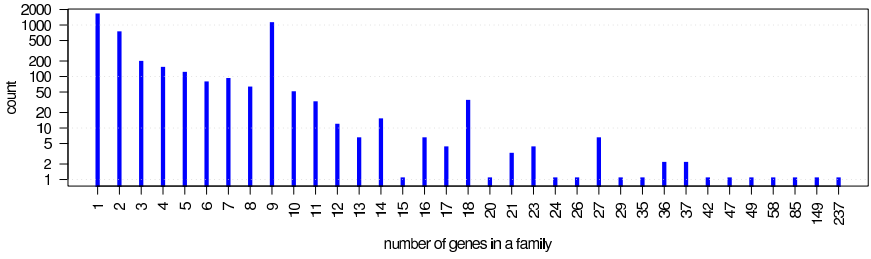


Fig. 1. 18573 coding genes have been clustered into 4072 families. For each observed family size a number of such families is shown. Single-gene families contain novel genes. Since nine genomes are analysed a peak for families containing a single gene for each genome is observed.

similar matrices are transformed into ones with lower similarity measure. Next, the shuffled matrices are compared using the same wmCompare algorithm. We repeat the shuffling 100 times and at the end we calculate the average number of pairs detected above the chosen threshold of matrices similarity. This average is a measure of the number of falsely discovered pairs of similar matrices in the original set.
We assume, that a typical transcription factor regulates more than a single gene. Therefore, we apply once again the MCL clustering algorithm to the graph constructed from the top pairs of similar matrices. This way we obtain groups of similar matrices which after alignment give us the profiles predicted to be recognized by a transcription factor. Finally, we perform a genome wide prediction of transcription factor binding sites using the obtained profiles. We use the approach proposed in [27] with parameters giving with probability 0.05 a single false positive binding site prediction for a sequence of length 500 nt.

2.2. *Experimental part*

Prochlorococcus sp. Med4 was grown in artificial seawater medium described previously [28] with a trace metal mix derived from medium Pro99 (Chisholm, personal communication). This modification resulted in the following final concentrations: 1.17 mM EDTA; 0.008 mM $ZnCl_2$; 0.005 mM $CoCl_2$; 0.09 mM $MnCl_2$; 0.003 mM Na_2MoO_4; 0.01 mM Na_2SeO_3; 0.01 mM $NiCl_2$; 1.17 mM $FeCl_3$. Cultures were kept under 10 μmol of photons $\cdot$ m^{-2} s^{-1} continuous blue light at $19 \pm 1°C$ and harvested by centrifugation at 10 200 g for 10 min in a Dupont RC5C centrifuge.
Total RNA was isolated as previously described [12]. Transcriptional initiation sites were determined by 5'-RACE following the method of [3] with modifications outlined in detail in [33].

3. Results

3.1. *Computational analysis*

Starting point for a phylogenetic footprinting analysis is the definition of the set of orthologous protein-coding genes between the genomes of interest. From all nine

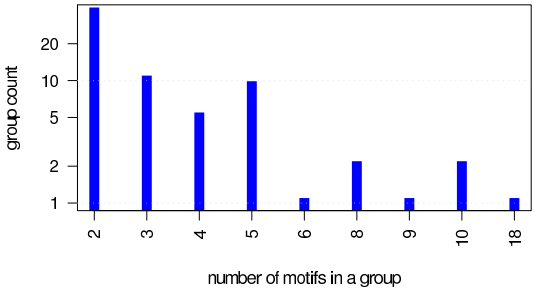
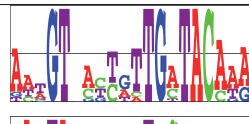
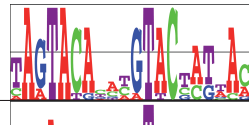
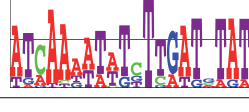


Fig. 2. Distribution of group sizes when 200 pairs of similar motifs were clustered.

genomes we could extract 18573 coding genes and after clustering we obtained 4072 orthologous gene families. Fig. 1 shows the distribution of gene families for different cluster sizes. We decided to study further approx. 35.1% of the clusters which contained at least six orthologs, to minimize problems resulting of small sample size when outcoming PSCMs are compared to each other.

For each gene family its set of gene upstream sequences was extracted and the best conserved motif in them was predicted. Then, all the obtained motifs were compared to each other. The outcoming list of motifs similarities was empirically limited to the top 200 pairs of motifs. Shuffling of motifs allowed us to estimate the average number of falsely discovered similar pairs in the top set of pairs to 132. Clusters of similar motifs were observed, suggesting existence of binding sites for *trans*-acting factors which control more than a single gene. Fig. 2 shows the distribution of number of motifs clustered into groups of high similarity. All 21 obtained clusters

Table 2. Predicted motifs and regulons identified in *Prochlorococcus* Med4 compared to all cyanobacteria motifs we could identify in literature. The ArsR motif can be assigned only manually since the most similar group contains less than four matrices.

name	known consensus	group size	predicted sequence logo	regulon
NtcA	TGT-N_{10}-ACA	10		*ntcA*; *spt,agt*; *glnA*; *glnB*; *urtA*
LexA	TAGTACA-N_2-TGTACTA	6		*recA*; *umuD*; *umuC*; *lexA*;
ArsR	ATCAA-N_6-TTGAT	2		*gap*; *arsR*; *pstS*; *phoB*; *phoR*

of elements having at least four motifs were merged and in Tab. 2 we list motifs similar to all previously known from literature. The received 21 clusters of simi-

lar motifs corresponded well with the number of expected biological motifs which was estimated from the number of annotated DNA-binding proteins within marine genomes (Tab. 1). Depending on the genome we observed 4 to 8 σ factors and 17 to 30 transcription factors which can be assumed to possess DNA-binding properties. Thus, a biological meaningful number of *cis* elements was suggested between 20 and 40 compared to the information of encoded genes. Finally, these motifs were used to search for candidate regulatory elements in upstream regions of all studied genomes. The computational analysis lasted seven hours on a single-CPU typical desktop computer.

The results of the genome-wide search were analyzed in detail by assigning the downstream genes to known pathways or regulons. Towards this goal, the genome annotation as well as KEGG[b] database and an intensive literature search were informative. This final evaluation revealed three motifs with analogy to already known sites for certain cyanobacteria: NtcA, LexA and ArsR, described in detail below. Moreover, we observed several of the predicted regulons belonging to riboswitch motifs (for example THI) or other non-protein binding elements [2], which were excluded from further investigations.

NtcA is a major regulator for nitrogen control in cyanobacterial cells [16]. Those parts of the genome, which are repressed or activated by its presence, constitute the N-regulon. Here, only a small but high-scoring subset of this putative regulon was defined, including genes for major enzymes of nitrogen-metabolizing pathways such as *spt,agt* (aminotransferase) and *glnA* (glutamine synthetase) as well as important nitrogen dependent transport systems like *urtABCDE* (urea transporter). The consensus sequence, identified here, harbors additional features besides the often used TGT-N_{10}-ACA motif: The flanking A/T-rich sequences and a conserved TG (or CA) dimer [32]. Thus, our more complex motif of marine cyanobacteria corresponds partly to the profile GTA-N_8-TAC suggested recently [30].

The putative **LexA** site found for marine cyanobacteria is highly similar to the previously described consensus sequences of gram-positive and freshwater cyanobacteria [22]. Furthermore, the LexA regulon predicted here contains several genes known to be active in the SOS response system such as *umuC* and *umuD* and especially *recA* and *lexA*. RecA and LexA represent the positive and negative regulator respectively, which might indicate a mechanism surprisingly similar to the SOS system best known from *E. coli* [34].

An **ArsR**-like consensus sequence is located within the spacer region of *arsR* and *gap*. However, the *arsBHC* operon that is involved in arsenic sensing and resistance in *Synechocystis* PCC 6803 [21] was not found within the marine genomes. Thus, the ArsR-like factor here may participate in the regulation of other genes and operons. Indeed, ArsR-like sites were predicted upstream of genes like *pstS*, *phoB* (two-component response regulator, phosphate) and *phoR*, thought to be regulated by the amount of phosphate in the cell. As there is also a regulator for phosphate,

[b] `http://www.genome.jp/kegg/`

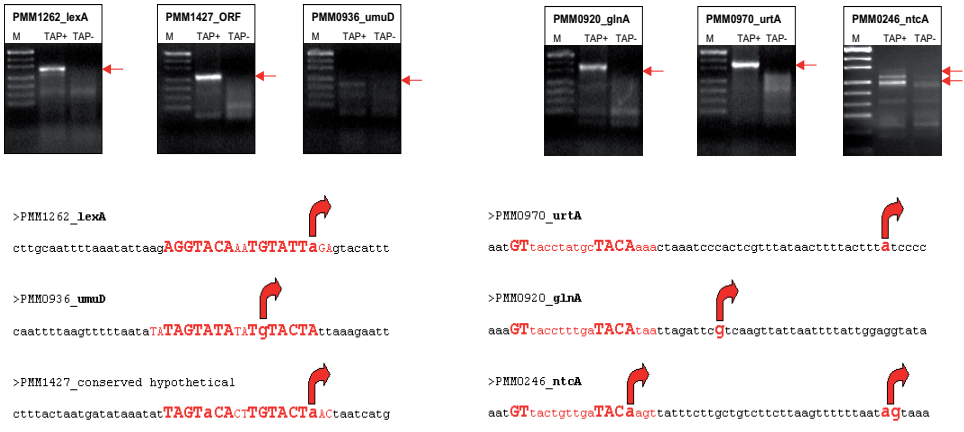


Fig. 3. Results of the PCR step during 5' RACE experiments for *lexA*, *umuD*, PMM1427 (left panel) and *urtA, glnA, ntcA* (right panel) in *Prochlorococcus* Med4. For each gene one single TIS appears except for *ntcA*, which exhibits two signals in the TAP-treated (TAP+) line. Overlay of the putative LexA recognition sequence (upper case letters) and the determined TIS (indicated by an arrow) for *lexA*, *umuD* and PMM1427 (left panel). Overlay of the predicted NtcA binding site and the mapped TIS upstream of *urtA*, *glnA*, *ntcA* (right panel).

PhoB, encoded in the marine genomes studied here, a crosstalk between both regulons might be assumed with the exception of AS9601 and MIT9515, where the ArsR-like regulator, encoded by the gene *arsR*, is missing.

3.2. *Experimental verification*

5' RACE experiments in *Prochlorococcus* Med4 were used to locate the transcription initiation site (TIS) of genes, for which putative DNA-binding sites had been predicted via our phylogenetic footprinting pipeline. For two of the best scoring motifs, NtcA and LexA, three genes were chosen, respectively. The TIS of *urtA*, *glnA*, *ntcA* as well as for *lexA*, *umuD* and PMM1427, a conserved hypothetical ORF, were mapped close to the predicted DNA-binding sites by RACE experiments. The results are shown in Fig. 3. All three predicted LexA motifs overlayed with the experimentally identified TIS as it might be assumed for LexA protein function as a repressor in bacterial gene transcription. The predicted motifs of NtcA showed different distances to the verified TIS: overlay, -10 as well as -35 distance was observed which can be easily explained by the dual function of NtcA as a repressor or activator for transcription.

4. Discussion

Phylogenetic footprinting was successfully applied to a set of sequenced marine genomes to reveal functionally relevant conservations between promoter regions of likely co-regulated genes. Thus, new information was obtained about the funda-

mentals of transcriptional regulation for marine cyanobacteria. In a first step, a set of orthologous coding regions was calculated resulting in 1428 families, which represents a number similar to other BLASTp comparisons of marine cyanobacterial genomes [8, 17]. Keeping in mind, that the total number of coding regions in these genomes varies between 1716 (Med4) and 2997 (MIT9303), at least around half of all genes belong to these conserved core gene families. Within this set of clusters, only 5 annotated sigma and 17 transcriptional factors were found, which likely constitute the core set of transcriptional regulatory proteins conserved between these nine marine cyanobacteria. Analyzing the orthologus upstream regions of family genes 21 motifs were detected above a chosen threshold which corresponded perfectly to an expected number of 20 to 40 DNA-binding sites. Motifs similar to previously described consensus sequences of the regulators NtcA and LexA known from freshwater cyanobacteria were identified as well as new regulatory motifs were predicted. Detailed analysis of six chosen promoters revealed that the predicted binding sites of LexA and NtcA belong to the experimentally defined promoter regions. Moreover, LexA is located exactly at the transcription initiation site for the studied genes including the *lexA* gene itself. Thus, LexA might be negatively autoregulated and could act as the repressor for several other genes. Although today, there are different functions for LexA discussed in literature [7, 14, 22] and studies about *Synechocystis* [7, 14] raised the question if all cyanobacteria possess an *E. coli*-type SOS regulon, the data obtained during this study of marine cyanobacteria give evidence for a DNA repair system surprisingly similar to the *E. coli* model.
Thus, a core set of regulons for the smallest known phototrophs is suggested here for the first time. The comparison of nine related genomes gives new insights into the minimum network of transcriptional regulation for strains within the marine ecosystem, but it does also allow drawing conclusions for cyanobacteria in general: Two known regulators, NtcA and LexA, appeared to be conserved over a wider evolutionary distance from freshwater to the group of marine cyanobacteria – from the most primitive unicellular to the filamentously growing complex species. The identification of NtcA and LexA in marine cyanobacteria illustrates how the data set might be utilized for an identification of promoters and regulatory sequences in other cyanobacterial species. In contrast, other factors like the one recognizing the ArsR-like binding site, might have evolved differentially and probably possess new functions and regulons adapted to the marine environment. Further experiments and comparisons with high throughput gene expression data will improve this initial regulatory network. Moreover, one has to remark that the computational predictions of DNA binding sites made here together with the experimentally tested examples can not serve as the entire proof of their biological function. For this purpose, additional binding studies, e.g. DNA affinity precipitation, DNase I protection or mobility shift assays, as well as detailed mutational analyses of the appropriate promoter regions might follow in the next future. Nevertheless, our global analysis represented here, could be a starting point to understand how these tiny and even so specialized organisms could dominate the oceans for millions of years although

environmental conditions were and are changing.

References

[1] Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J., Basic local alignment search tool, *J. Mol. Biol.*, 215(3):403–410, 1990.

[2] Axmann, I.M., Kensche, P., Vogel, J., Kohl, S., Herzel, H., and Hess, WR., Identification of cyanobacterial non-coding RNAs by comparative genome analysis, *Genome Biol.*, 6(9):R73, 2005.

[3] Bensing, B.A., Meyer, B.J., and Dunny, G.M., Sensitive detection of bacterial transcription initiation sites and differentiation from RNA processing sites in the pheromone-induced plasmid transfer system of Enterococcus faecalis, *Proc. Natl. Acad. Sci. USA*, 93(15):7794–7799, 1996.

[4] Bulyk, M.L., Computational prediction of transcription-factor binding site locations, *Genome Biol.*, 5(1):201, 2003.

[5] Cliften, P.F., Hillier, L.W., Fulton, L., Graves, T., Miner, T., Gish, W.R., Waterston, R.H., and Johnston, M., Surveying Saccharomyces genomes to identify functional elements by comparative DNA sequence analysis, *Genome Res.*, 11(7):1175–1186, 2001.

[6] Dieterich, C., Grossmann, S., Tanzer, A., Ropcke, S., Arndt, P.F., Stadler, P.F., and Vingron, M., Comparative promoter region analysis powered by CORG, *BMC Genomics*, 6(1):24, 2005. Comparative Study.

[7] Domain, F., Houot, L., Chauvat, F., and Cassier-Chauvat, C., Function and regulation of the cyanobacterial genes lexA, recA and ruvB: LexA is critical to the survival of cells facing inorganic carbon starvation, *Mol. Microbiol*, 53(1):65–80, 2004.

[8] Dufresne, A., Garczarek, L., and Partensky, F., Accelerated evolution associated with genome reduction in a free-living prokaryote, *Genome Biol.*, 6(2):R14, 2005.

[9] Enright, A.J., Van Dongen, S., and Ouzounis, C.A., An efficient algorithm for large-scale detection of protein families, *Nucleic Acids Res.*, 30(7):1575–1584, 2002.

[10] Frazer, K.A., Elnitski, L., Church, D.M., Dubchak, I., and Hardison, R.C., Cross-species sequence comparisons: a review of methods and available resources, *Genome Res.*, 13(1):1–12, 2003.

[11] Frith, M.C., Hansen, U., Spouge, J.L., and Weng, Z., Finding functional sequence elements by multiple local alignment, *Nucleic Acids Res.*, 32(1):189–200, 2004.

[12] Garcia-Fernandez, J.M., Hess, W.R., Houmard, J., and Partensky, F., Expression of the psbA gene in the marine oxyphotobacteria Prochlorococcus spp, *Arch Biochem Biophys*, 359(1):17–23, 1998.

[13] Gottgens, B., Gilbert, J.G., Barton, L.M., Grafham, D., Rogers, J., Bentley, D.R., and Green, A.R., Long-range comparison of human and mouse SCL loci: localized regions of sensitivity to restriction endonucleases correspond precisely with peaks of conserved noncoding sequences, *Genome Res.*, 11(1):87–97, 2001.

[14] Gutekunst, K., Phunpruch, S., Schwarz, C., Schuchardt, S., Schulz-Friedrich, R., and Appel, J., LexA regulates the bidirectional hydrogenase in the cyanobacterium Synechocystis sp. PCC 6803 as a transcription activator, *Mol. Microbiol*, 58(3):810–823, 2005.

[15] Hagstrom, A., Pommier, T., Rohwer, F., Simu, K., Stolte, W., Svensson, D., and Zweifel, U.L., Use of 16S ribosomal DNA for delineation of marine bacterioplankton species, *Appl. Environ Microbiol*, 68(7):3628–3633, 2002.

[16] Herrero, A., Muro-Pastor, A.M., and Flores, E., Nitrogen control in cyanobacteria, *J. Bacteriol*, 183(2):411–425, 2001.

[17] Hess, W.R., Genome analysis of marine photosynthetic microbes and their global role, *Curr. Opin. Biotechnol*, 15(3):191–198, 2004.
[18] Kielbasa, S., The BioMinerva framework (in preparation), 2007.
[19] Kielbasa, S.M., Gonze, D., and Herzel, H., Measuring similarities between transcription factor binding sites, *BMC Bioinformatics*, 6:237, 2005.
[20] Loots, G.G., Locksley, R.M., Blankespoor, C.M., Wang, Z.E., Miller, W., Rubin, E.M., and Frazer, K.A., Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons, *Science*, 288(5463):136–140, 2000.
[21] Lopez-Maury, L., Florencio, F.J., and Reyes, J.C., Arsenic sensing and resistance system in the cyanobacterium Synechocystis sp. strain PCC 6803, *J. Bacteriol*, 185(18):5363–5371, 2003.
[22] Mazon, G., Lucena, J.M., Campoy, S., Fernandez de Henestrosa, A.R., Candau, P., and Barbe, J., LexA-binding sequences in Gram-positive and cyanobacteria are closely related, *Mol. Genet. Genomics*, 271(1):40–49, 2004.
[23] McGuire, A.M., Hughes, J.D., and Church, G.M., Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes, *Genome Res.*, 10(6):744–757, 2000.
[24] Moore, L.R., Rocap, G., and Chisholm, S.W., Physiology and molecular phylogeny of coexisting Prochlorococcus ecotypes, *Nature*, 393(6684):464–467, 1998.
[25] Palinska, K.A., Jahns, T., Rippka, R.,and Tandeau De Marsac, N., Prochlorococcus marinus strain PCC 9511, a picoplanktonic cyanobacterium, synthesizes the smallest urease, *Microbiology*, 146 Pt 12:3099–3107, 2000.
[26] Partensky, F., Hess, W.R., and Vaulot, D., Prochlorococcus, a marine photosynthetic prokaryote of global significance, *Microbiol Mol. Biol. Rev.*, 63(1):106–127, 1999.
[27] Rahmann, S., Muller, T., and Vingron, M., On the power of profiles for transcription factor binding site detection, *Stat. Appl. Genet. Mol. Biol.*, 2:Article7, 2003.
[28] Rippka, R., Coursin, T., Hess, W., Lichtle, C., Scanlan, D.J., Palinska, K.A., Iteman, I., Partensky, F., Houmard, J., and Herdman, M., Prochlorococcus marinus Chisholm et al. 1992 subsp. pastoris subsp. nov. strain PCC 9511, the first axenic chlorophyll a2/b2-containing cyanobacterium (Oxyphotobacteria), *Int. J. Syst. Evol. Microbiol*, 50 Pt 5:1833–1847, 2000.
[29] Su, Z., Dam, P., Chen, X., Olman, V., Jiang, T., Palenik, B., and Xu, Y., Computational inference of regulatory pathways in microbes: an application to phosphorus assimilation pathways in Synechococcus sp. WH8102, *Genome Inform.*, 14:3–13, 2003.
[30] Su, Z., Olman, V., Mao, F., and Xu, Y., Comparative genomics analysis of NtcA regulons in cyanobacteria: regulation of nitrogen assimilation and its coupling to photosynthesis, *Nucleic Acids Res.*, 33(16):5156–5171, 2005.
[31] Ureta-Vidal, A., Ettwiller, L., and Birney, E., Comparative genomics: genome-wide analysis in metazoan eukaryotes, *Nat. Rev. Genet.*, 4(4):251–262, 2003.
[32] Vazquez-Bermudez, M.F., Flores, E., and Herrero, A., Analysis of binding sites for the nitrogen-control transcription factor NtcA in the promoters of Synechococcus nitrogen-regulated genes, *Biochim. Biophys. Acta*, 1578(1-3):95–98, 2002.
[33] Vogel, J., Axmann, I.M., Herzel, H., and Hess, W.R., Experimental and computational analysis of transcriptional start sites in the cyanobacterium Prochlorococcus MED4, *Nucleic Acids Res.*, 31(11):2890–2899, 2003.
[34] Walker, G.C., Mutagenesis and inducible responses to deoxyribonucleic acid damage in Escherichia coli, *Microbiol Rev.*, 48(1):60–93, 1984.
[35] Wasserman, W.W. and Sandelin, A., Applied bioinformatics for the identification of regulatory elements, *Nat. Rev. Genet.*, 5(4):276–287, 2004.
[36] West, N.J. and Scanlan, D.J., Niche-partitioning of Prochlorococcus populations in a

stratified water column in the eastern North Atlantic Ocean, *Appl. Environ Microbiol*, 65(6):2585–2591, 1999.

[37] Yan, B., Methe, B.A., Lovley, D.R., and Krushkal, J., Computational prediction of conserved operons and phylogenetic footprinting of transcription regulatory elements in the metal-reducing bacterial family Geobacteraceae, *J. Theor. Biol.*, 230(1):133–144, 2004.