Combining microarray and genomic data to predict DNA binding motifs

Linyong Mao,¹ Chris Mackenzie,² Jung H. Roh,² Jesus M. Eraso,² Samuel Kaplan² and Haluk Resat¹

Correspondence

Haluk Resat

haluk.resat@pnl.gov

¹Pacific Northwest National Laboratory, Computational Biology and Bioinformatics Group, PO Box 999, MS: K7-90, Richland, WA 99352, USA

²Department of Microbiology and Molecular Genetics, The University of Texas Health Science Center, Medical School, Houston, TX 77030, USA

The ability to detect regulatory elements within genome sequences is important in understanding how gene expression is controlled in biological systems. In this work, microarray data analysis is combined with genome sequence analysis to predict DNA sequences in the photosynthetic bacterium *Rhodobacter sphaeroides* that bind the regulators PrrA, PpsR and FnrL. These predictions were made by using hierarchical clustering to detect genes that share similar expression patterns. The DNA sequences upstream of these genes were then searched for possible transcription factor recognition motifs that may be involved in their co-regulation. The approach used promises to be widely applicable for the prediction of *cis*-acting DNA binding elements. Using this method the authors were independently able to detect and extend the previously described consensus sequences that have been suggested to bind FnrL and PpsR. In addition, sequences that may be recognized by the global regulator PrrA were predicted. The results support the earlier suggestions that the DNA binding sequence of PrrA may have a variable-sized gap between its conserved block elements. Using the predicted DNA binding sequences, a whole-genome-scale analysis was performed to determine the relative importance of the interplay between the three regulators PpsR, FnrL and PrrA. Results of this analysis showed that, compared to the regulation by PpsR and FnrL, a much larger number of genes are candidates to be regulated by PrrA. The study demonstrates by example that integration of multiple data types can be a powerful approach for inferring transcriptional regulatory patterns in microbial systems, and it allowed the detection of photosynthesis-related regulatory patterns in *R. sphaeroides*.

Received 29 April 2005

Revised 25 July 2005

Accepted 26 July 2005

INTRODUCTION

The purple non-sulfur photosynthetic bacterium *Rhodobacter sphaeroides* 2.4.1 is well known for its remarkable metabolic versatility. It is capable of growing aerobically, anaerobically (in the dark in the presence of external electron acceptors such as DMSO), photosynthetically in the light without oxygen, and fermentatively. To adapt to environmental changes, gene expression is controlled by a hierarchy of regulatory elements. For example, the expression of the photosynthesis genes of *R. sphaeroides* is primarily controlled through the interplay of three major regulatory systems, the PrrB/PrrA two-component system (Eraso & Kaplan, 1994; Lee & Kaplan, 1992), the AppA/PpsR antirepressor/repressor system (Gomelsky & Kaplan, 1997), and the FnrL regulator (Oh & Kaplan, 2001; Zeilstra-Ryalls & Kaplan, 1995).

In the PrrB/PrrA (photosynthetic response regulator) two-component system, PrrA serves as a response regulator, and PrrB (Lee & Kaplan, 1992) is a membrane-localized sensor kinase/phosphatase which phosphorylates PrrA upon O₂ deprivation (Eraso & Kaplan, 1994). In addition to regulating photosynthesis-gene expression, PrrA acts as a global regulator, affecting the expression of genes encoding electron-transport components, genes involved in CO₂ and N₂ fixation, and genes involved in hydrogen oxidation, among others (Elsen *et al.*, 2004; Joshi & Tabita, 1996; Qian & Tabita, 1996). Although the importance of the role played by PrrA in gene regulation is clear, the DNA sequence to which it binds remains poorly defined.

In the AppA/PpsR antirepressor/repressor system (Gomelsky & Kaplan, 1997), AppA (activation of photopigment and *puc* expression) serves as an antirepressor and modulates the repressor activity of PpsR (photopigment suppression) (Penfold & Pemberton, 1994) such that PpsR becomes more active upon the oxidation of the quinone pool (Braatsch *et al.*, 2002; Oh & Kaplan, 2001). The antirepressor AppA is

Abbreviation: CHPC, cluster with high photosynthesis content.

Supplementary tables and figures are available with the online version of this paper.

also responsible for blue-light photoreception, which can affect its activity toward PpsR (Braatsch *et al.*, 2002; Masuda & Bauer, 2002). PpsR functions as a tetramer with a helix–turn–helix (HTH) domain at the carboxy-terminal region that genetic analysis suggests binds to a conserved DNA sequence, TGTN₁₂ACA, where N represents a non-specific nucleotide (Gomelsky *et al.*, 2000). This DNA motif is found in the region upstream of the genes *bch* and *crt*, as well as the *puc* operon, all of which encode products required for photosynthesis, i.e. bacteriochlorophyll, carotenoids and structural proteins, respectively (Zeilstra-Ryalls *et al.*, 1998).

The *R. sphaeroides* regulator FnrL is considered to be a homologue of the *Escherichia coli* anaerobic regulatory protein FNR (fumarate and nitrate reduction regulatory protein) (Zeilstra-Ryalls & Kaplan, 1995). This hypothesis is based in part on the FnrL amino acid sequence, which shows homology to known functional domains of the FNR protein. By analogy, it has also been hypothesized that FnrL may recognize the FNR consensus sequence TTGATN₄ATCAA (Zeilstra-Ryalls & Kaplan, 1998). This consensus sequence has been found in the sequences upstream of *hemA*, *hemN* and *hemZ*, genes involved in the tetrapyrrole biosynthetic pathway, the *bchE* gene, and the *puc* operon (Choudhary & Kaplan, 2000; Zeilstra-Ryalls & Kaplan, 1995). Regions upstream of the *ccoNOQP* operon encoding the *cbh₃* oxidase, the *rdxBHIS* operon, and the structural genes encoding the *aa₃* cytochrome oxidase (Zeilstra-Ryalls *et al.*, 1998) also contain the FnrL consensus sequence, suggesting that FnrL indirectly regulates the volume of electron flow toward different terminal oxidases and to the Rdx redox centre by changing their gene-expression levels.

The purpose of this study was to predict and identify DNA motifs present in the *R. sphaeroides* 2.4.1 genome that may bind the transcription factors PrrA, PpsR and FnrL, and thereby identify which genes in the genome may be influenced by these regulators. The rationale behind our methodological approach was as follows: 'If genes a, b, c, d and e, show high levels of expression under condition x, and low levels under condition y, and no expression under condition z, then it is plausible that the expression of these genes may be controlled by the same regulatory protein. If so, then this regulator is hypothesized to recognize the same signature within the DNA sequence.' In brief, we carried out hierarchical clustering of *R. sphaeroides* genes using microarray mRNA expression data to follow which genes showed concomitant increased/decreased expression patterns under seven different experimental conditions. We then searched loci, i.e. the regions upstream of these genes or their operons, for signature sites that suggest co-regulation. These sites were then used to generate a predicted consensus sequence.

The application of both microarray data clustering and motif-finding approaches to a large dataset has allowed us to independently find putative PpsR binding sites that are in good agreement with the previously published PpsR binding consensus. It has also allowed us to predict refinements to

that consensus. Our results for FnrL binding sites are also in agreement with the previous predictions for the FnrL consensus sequence, although here we extend the likely numbers of target genes. For PrrA, our predictions suggest a PrrA DNA binding sequence comprising two blocks with an internal gap of variable length, again consistent with previously published predictions. We have also calculated the statistical distribution of the variable gap widths between the conserved block elements of the binding motif for PrrA. Using the predicted PpsR, FnrL and PrrA consensus sequences deduced from this study, we were able to predict the genes that are potentially regulated by these transcription factors throughout the genome. We note that due to the statistical filtering approach which was used and the limited amount of data available, our findings are likely to contain false-positive and false-negative binding sites. However, our analysis of microarray data from a PrrA mutant suggests that our method is sufficiently robust to assist in the prediction of genes controlled by this regulator. These newly identified target genes and their mode of regulation are now more amenable to study using classic genetic and biochemical approaches, in other words our findings will be used to design new experiments for the next round of studies.

METHODS

For clustering analysis, *R. sphaeroides* 2.4.1 (wild-type) and two mutant lines were examined. The wild-type was grown under five different growth conditions and the two mutant strains under one growth condition, a total of seven experiments. All strains, described in detail below, were grown as independent triplicate cultures (Roh *et al.*, 2004). RNA was harvested from each culture, converted to cDNA, then applied to its own microarray chip. The findings described in the Results are derived from the data obtained from 21 independent microarray experiments (seven conditions with triple repeats).

For validation of the methodology, *R. sphaeroides* 2.4.1 and a *prpA*[−] mutant (PRRA2) were grown under anaerobic dark DMSO conditions. Both strains were grown as independent triplicate cultures and treated as described above.

Strains. *R. sphaeroides* strain 2.4.1 (wild-type, ATCC BAA-808) and two in-frame deletion mutation-containing strains, *ccoNOQP* (Oh & Kaplan, 2002) and *rdxB* (Oh & Kaplan, 1999), were used in this study. The mutant strains are defective in part of the known signal-transduction pathway for photosynthesis gene expression. In the wild-type, the photosynthesis genes are only expressed under anaerobic conditions; however, in these two mutant strains, these genes are expressed under aerobic conditions. We included the mutant data in our analysis because in terms of a statistical approach, any data pertaining to photosynthesis gene expression can add to the information content by providing a larger dataset to further enhance the analysis.

The *prpA*[−] mutation used for validation was created by deletion of part of the *prpA* gene and has been described previously (Eraso & Kaplan, 1997).

***R. sphaeroides* growth conditions.** Briefly, wild-type cells were grown under the following five growth conditions: aerobic (30 % O₂), photosynthetic (3, 10 and 100 W m^{−2} light intensity), and

DMSO with 10 W m⁻² light intensity. The two mutant strains were only grown under aerobic, i.e. 30 % O₂ conditions. For validation of the study, both wild-type and PRRA2 were grown under anaerobic dark DMSO conditions.

In detail, the strains were grown at 29 ± 1 °C on Sistrom's minimal medium A (SIS) containing succinate as carbon source (Sistrom, 1962). Aerobic cultures were grown while sparging with a gas mixture of 30 % O₂/69 % N₂/1 % CO₂ and harvested at a low OD₆₀₀ of 0.18 ± 0.02 in order to ensure oxygen saturation. Photosynthetic cultures were grown at light intensities of 3, 10 and 100 W m⁻² (measured at the surface of the growth vessel) while sparging with 95 % N₂/5 % CO₂, and harvested at OD₆₀₀ 0.45 ± 0.05 to prevent self-shading. For cultures grown with DMSO at 10 W m⁻², the cells were cultivated in the presence of 60 mM DMSO (to change the redox state of the cells), and were also sparged with a gas mixture of 95 % N₂/5 % CO₂ (to generate anaerobic conditions) and harvested at OD₆₀₀ 0.45 ± 0.05. All light intensities were measured using a YSI-Kettering model 65A radiometer (Simpson Electric Co.).

For validation, wild-type and PRRA2 were grown in Sistrom's medium containing a final concentration of 60 mM DMSO. The medium was sparged with 95 % N₂/5 % CO₂. Cells were harvested at the densities described above.

RNA manipulation. A previously described RNA isolation procedure (Roh & Kaplan, 2002) was modified to optimize the isolation of intact mRNA for microarray analysis (Roh *et al.*, 2004). We modified the earlier procedure by eliminating cell collection by centrifugation. A volume of cells grown as described above was directly pipetted into an equal volume of 2 × lysis buffer (100 °C). After thorough mixing, lysed cells were immediately transferred to an equal volume of hot phenol solution (65 °C). The total time required to transfer from the culture vessel to hot phenol was kept to less than 1 min to minimize mRNA degradation and to maximize the yield of intact mRNA. The remainder of the RNA purification procedure was identical to that described previously (Roh & Kaplan, 2002). Each isolated RNA sample was treated with 50 µl RQ1 RNase-free DNase (1 unit µl⁻¹, Promega) and 50 µl 10 × buffer in a total volume of 500 µl. Samples were incubated for 1 h at 37 °C, extracted with acidic phenol, acidic phenol/chloroform, and chloroform, then precipitated by adding 1 ml ethanol. The pellet was washed with 75 % ethanol and suspended in diethylpyrocarbonate (DEPC)-treated water. Total RNA was pelleted again by adding the same volume of 4 M LiCl, washed with 75 % ethanol, and resuspended in DEPC-treated water. Chromosomal DNA contamination was tested by PCR amplification using the *rdxB*-specific primers (*a* and *b*), as described previously (Roh & Kaplan, 2002).

Microarray experiments. The *R. sphaeroides* 2.4.1 GeneChip was custom designed and manufactured by Affymetrix Inc. (Pappas *et al.*, 2004). In most cases, one probe set was designed to represent one gene/ORF. But there are cases where more than one probe set with the same RSP number (e.g. *RSP1556_f_at* and *RSP1556_r_at*) was used to represent the same gene/ORF. Total RNA was prepared from three independent cultures of *R. sphaeroides*. cDNA synthesis, fragmentation, labelling and hybridization were adapted, with few modifications, from the methods optimized for the GeneChip designed for the *Pseudomonas aeruginosa* Genome Array by Affymetrix, Inc. (<http://www.affymetrix.com/support/technical/manuals.affx>). Briefly, 10 µg total RNA was annealed with 750 ng of random primers (New England Biolabs) and incubated at 70 °C for 10 min, and then at 25 °C for 1 h. First-strand cDNA was synthesized with 200 units µl⁻¹ SuperScript II with 5 × 1st strand buffer (Invitrogen Life Technologies) in the presence of 10 mM DTT, 0.5 mM dNTPs and 0.5 units µl⁻¹ SUPERase In RNase inhibitor (Ambion) (25 °C for 20 min, 37 °C for 1 h, 42 °C for 10 min). After removal of RNA by alkaline treatment and neutralization, the cDNA

synthesis product was purified using the QIAquick PCR purification kit (Qiagen). For fragmentation, 7–9 µg cDNA and 1 unit of RQ1 DNase I (Promega) were incubated at 37 °C. After 1 min, one-third of the cDNA/DNase mixture was removed and heat-inactivated at 100 °C for 5 min. Further one-third aliquots were removed at 2 and 3 min and similarly heat-inactivated. The desired cDNA size range of 50–200 bases was selected after 3 % agarose gel electrophoresis using 200 ng fragmented cDNA. The fragmented cDNA was 3'-end labelled using the Enzo BioArray Terminal Labelling kit (Affymetrix) with biotin-ddUTP. Target hybridization, washing, staining and scanning were performed according to the protocol supplied by the manufacturer using a GeneChip Hybridization Oven 640, a Fluidics Station 400, and the Agilent GeneArray Scanner under the control of Affymetrix Microarray Suite 5.0.

Data files were analysed using the MAS 5.0 (Affymetrix Inc.) and dChip 1.2 software (Li & Hung Wong, 2001; Li & Wong, 2001). Raw intensity values from different experiments were normalized against a target intensity value for across-experiment comparison. Probe intensities of the triplicate array experiments for every condition were then further intensity-normalized using the total array intensity of the chips. The mean of triplicate measurements was used to describe the expression level of a gene for that particular condition, and the mean expression values for the seven experimental conditions were then used in the clustering analysis.

Clustering analysis. Genes were clustered according to their expression patterns in the seven different experiments using the dChip software (Li & Hung Wong, 2001; Li & Wong, 2001). The hierarchical clustering method used within dChip has been described elsewhere (Eisen *et al.*, 1998). Before clustering, genes that showed a relative expression variation (ratio of the standard deviation to mean value) of less than 0.5 over the seven experiments were determined in order to filter out the genes whose expression change across the studied conditions was insignificant. Of the original 4490 probe sets, 3583 fell into this class and were removed from further analysis. The remaining 907 probe sets that showed significant changes were used in the clustering analysis. It should be noted that the cutoff used in the selection is somewhat arbitrary, and this approach has the potential to weigh towards genes expressed at low levels. To verify that our filtering approach does not introduce a serious selection bias, we have calculated the distribution of the intensities of the genes for the total and the selected sets. Computed distributions (Supplementary Fig. S1) clearly showed that the filtering schema utilized does not cause a noticeably significant statistical bias.

The expression values of the 907 probe sets used in the clustering analysis were further preprocessed such that they had a zero mean and unit standard deviation over the seven experiments. The analysis utilized the average linkage method, in which the distance between pairs of genes is defined as 1 – R, where R is the correlation coefficient between the expression patterns.

DNA motif search. The MEME (Bailey & Elkan, 1994) and BioProspector (Liu *et al.*, 2001) programs were used to search the DNA sequences upstream of genes for DNA binding motifs. In this work, we often refer to the sequences upstream of genes and operons collectively as loci. We use this term for convenience, but the reader should realize that it can mean that the sequence originated from upstream of a gene or an operon. Up to 1 kb of sequence upstream from an individual gene or from the first gene in each operon was extracted from the genomic sequence and used in these motif searches. When available, the structures of operons were obtained from the literature (Oh & Kaplan, 2001); otherwise they were predicted based on the relative chromosomal positions of the genes, their putative transcription directions, their intergenic sequence lengths or their functions.

Given a group of related DNA or protein sequences, the MEME program (Bailey & Elkan, 1994) uses a statistical expectation maximization technique to find different fixed-width motifs. The BioProspector program (Liu *et al.*, 2001) uses a Gibbs sampling strategy to detect sequence motifs, and the motifs can be allowed to have variable widths. Using the putative DNA binding motifs detected by MEME and BioProspector, the MAST program (Bailey & Gribskov, 1998) was then applied to scan sequences upstream of the target genes to search for matches to the detected motifs. For our analysis, we modified the MAST program so that it could search for motifs with variable widths.

The two chromosomes of *R. sphaeroides* are predicted to encode about 3980 genes, 2095 (53 %) of which have intergenic upstream sequences (loci) with lengths ≥ 50 bp. We collected the intergenic upstream sequences of these 2095 genes and, in addition, we collected the sequence upstream of *pucB* (2096 upstream sequences in total). This latter sequence was dealt with separately because its 5' end overlaps with the 3' end of an upstream hypothetical gene (RSP0313). This gene organization, i.e. the lack of an intergenic sequence, would have prevented the *pucB* upstream sequence from being captured by the ≥ 50 bp cutoff described above. It was known that the *pucB* promoter is embedded in the coding sequence of the upstream gene. In addition, *pucB* has been shown experimentally to be regulated by PpsR/FnrL/PrrA (Lee & Kaplan, 1992); it was therefore important to deal with it as a special case and thus include it in the analysis.

The *R. sphaeroides* genome sequence, the chromosomal locations of the encoded genes, their upstream sequences and annotations can be accessed at the website <http://genome.ornl.gov/microbial/rsph/>.

Consensus diagrams (cf. Figs 2 and 3) were created using the WebLogo program (Crooks *et al.*, 2004) through the website at <http://weblogo.berkeley.edu/>.

RESULTS

Cluster analysis of gene expression patterns

Clustering analysis allowed categorization of the genes according to the expression patterns that they exhibit. Genes belonging to the same cluster may be involved in functionally related biological activities and possibly be regulated through similar mechanisms. Our clustering analysis included seven different experimental conditions. As detailed in Methods, after filtering out genes that showed insignificant changes in expression level between different conditions, 907 probe sets remained that were subject to the subsequent hierarchical clustering.

Chromosome I of *R. sphaeroides* contains a contiguous ~ 67 kb region that encompasses the photosynthesis gene cluster and encodes the *puc* and *puf* operons, *bch* genes, *crt* genes, photosynthesis gene regulators and other photosynthesis-related genes (Choudhary & Kaplan, 2000). Seventy-nine of the probe sets on the microarray chip represent the genes located in this 67 kb photosynthesis region. Thirty-seven of these were included among the 907 probe elements selected for clustering, and constituted 4.1 % of the probe elements.

Two of the clusters generated by clustering analysis contained a significant number of genes and operons that lie within the 67 kb region of chromosome I and are

functionally integrated with photosynthesis. These clusters will be referred to as clusters with high photosynthesis content (CHPC) (see Fig. 1). CHPC1 is composed of 65 probe sets, of which 21 (32 %) lie within the 67 kb photosynthesis gene region. CHPC2 is composed of 44 probe sets, of which 10 (23 %) lie within the 67 kb photosynthesis region. Probe sets relating to photosynthesis that are contained within these two CHPCs are listed in Table 1. In our analysis, 38 and 23 loci were derived from CHPC1 and CHPC2, respectively. Since four loci were common to CHPC1 and CHPC2, the combined clusters contained 57 loci. The complete list of genes and operons contained in the two clusters is included in Supplementary Table S1.

The expression patterns of genes in CHPC1 are similar to those in CHPC2. The only notable difference between these two clusters is that when compared to aerobic growth conditions, most genes in CHPC1 increase in expression under photosynthetic conditions, i.e. 100, 10 and 3 W m⁻², and exhibit their highest levels of expression under the growth condition of 10 W m⁻² with DMSO (Fig. 1A), whereas genes in the CHPC2 cluster (Fig. 1B) exhibit their highest expression levels under 10 W m⁻² photosynthetic growth in the absence of DMSO.

DNA binding motif search

Since CHPCs contain a large number of genes functionally related to photosynthesis, investigating the regulation of the genes belonging to these clusters can help us understand the transcriptional regulation of photosynthesis gene expression in *R. sphaeroides*. The loci within the CHPCs were searched using the MEME and BioProspector programs. We first searched for motifs in the loci of each individual CHPC and then repeated the searches after combining the clusters. We first present the motifs detected in the loci and then discuss the properties of the predicted DNA recognition motifs that were detected. We particularly emphasize the detection of the putative PrrA DNA binding motif specific to *R. sphaeroides*.

CHPC1. The MEME program was used to search and generate the six most statistically significant motifs in the loci belonging to CHPC1. To be inclusive, a window of 6–50 bp for the motif length was used in the search. Each upstream sequence was allowed to contain any number of occurrences of each detected motif. Among the top six detected motifs, three were found to be of particular interest.

The first detected motif, TGTC[A/G][C/A]NNAANTT-GACA, has a 6 bp inverted repeat form and reproduces the known less-restrictive DNA binding sequence pattern that has been suggested to be recognized by PpsR: TGTT₁₂ACA (Gomelsky *et al.*, 2000). This was the highest-ranking motif in the search; its probability score matrix is reported in Supplementary Table S2. The second motif (ranked second; Supplementary Table S3), TTGA[T/C][C/A][G/A/T][G/C][A/G]TCAA, also has a palindromic structure and matches the hypothesized FnrL consensus sequence TTGATN₄ATCAA

Fig. 1. Hierarchical relationship diagrams of the two clusters that contain a large proportion of photosynthesis genes: (A) CHPC1 and (B) CHPC2. Each column corresponds to the following seven experiments (from left to right): aerobic, wild-type; aerobic, *ccoNOQP* mutant; aerobic, *rdxB* mutant; 100 W m⁻² without O₂ (photosynthetic, high light), wild-type; 10 W m⁻² without O₂ (photosynthetic, medium light), wild-type; 3 W m⁻² without O₂ (photosynthetic, low light), wild-type; and 10 W m⁻² without O₂ in the presence of DMSO, wild-type. Each row corresponds to an individual probe set. Expression values of each probe set are standardized to have a zero mean and unit standard deviation over the seven experimental conditions. Dark blue represents low expression levels and dark red, high expression levels. The corresponding colour bar gives the standardized expression values. Probe sets with a blue font colour represent genes in the 67 kb photosynthesis gene region of the chromosome. Colour-coded boxes before the gene names identify the functional families of genes, according to COG classification. The blue vertical line between the clustering diagram and COG boxes represents the range of the cluster, and a small horizontal line intersecting the vertical line marks the root node of the cluster.

(Zeilstra-Ryalls & Kaplan, 1998). We note that the detected PpsR and FnrL DNA binding motifs have perfect inverted-repeat forms, even though a palindromic structure was not

imposed in the search. The third detected motif (ranked sixth), GC[G/T][G/T/C]C[C/A/T]C[T/G]CT[G/T]CC[G/T]C, has a 5 bp inverted-repeat region that is poorly conserved

Table 1. Probe sets of the 67 kb photosynthesis region contained in the CHPC clusters.

Probe set*	Gene name	Gene function
(A) CHPC1		
RSP0255	<i>pufX</i>	Facilitates light-driven cyclic electron transfer
RSP0256	<i>pufM</i>	Photosynthetic reaction centre M subunit
RSP0257	<i>pufL</i>	Photosynthetic reaction centre L subunit
RSP0259	<i>Q</i>	Involved in spectral complex assembly
RSP0265	<i>crtE</i>	Geranylgeranyl pyrophosphate synthase
RSP0267	<i>crtC</i>	Hydroxyneurosporene dehydrogenase
RSP0268	N/A†	Unknown
RSP0270	<i>crtB</i>	Prephytoene pyrophosphate synthase
RSP0271	<i>crtI</i>	Phytoene dehydrogenase
RSP0272	<i>crtA</i>	Spheroidene monooxygenase
RSP0273	<i>bchI</i>	Protoporphyrin IX chelatase I subunit
RSP0274	<i>bchD</i>	Protoporphyrin IX chelatase D subunit
RSP0277	<i>bchP</i>	Geranylgeranyl hydrogenase
RSP0278	N/A	Hypothetical protein
RSP0279	<i>bchG</i>	Chlorophyll synthase 33 kDa subunit
RSP0280	<i>bchJ</i>	4-Vinyl reductase
RSP0281	<i>bchE</i>	Protoporphyrin IX monomethyl ester oxidative cyclase subunit
RSP0286	<i>bchB</i>	Protochlorophyllide reductase subunit
RSP0290	N/A	Hypothetical protein
RSP0315	<i>pucC</i>	Light-harvesting complex assembly
RSP0317	<i>hemN</i>	Coproporphyrinogen III oxidase, anaerobic
(B) CHPC2		
RSP0269	<i>tspO</i>	Outer-membrane protein, pore
RSP0276	<i>idi</i>	Isopentenyl-diphosphate isomerase
RSP0283	<i>ppaA</i>	Regulatory protein
RSP0291	<i>puhA</i>	Photosynthetic reaction centre H protein
RSP0292	N/A	Hypothetical membrane protein
RSP0314‡	<i>pucB</i>	Light-harvesting complex II β subunit

*Multiple genes grouped together represent genes in the same operon unit.
†N/A, unknown or hypothetical protein.
‡RSP0314, the *pucB* gene, is represented by five identical probes on the chip.

and resembles DNA recognition sequences that have previously been proposed for PrrA (Supplementary Table S4). The identification and possible biological significance of this predicted highly degenerate PrrA binding motif will be discussed in depth later.

CHPC2. MEME searching parameters for CHPC2 were identical to those for CHPC1, except that the length of the motif was set more stringently to vary between 14 and 20 bp to limit the lengths of detected motifs. Supplementary Tables S2–S4 list the three top-ranked motifs

detected by MEME. Comparison of the motif results for CHPC2 with the results for CHPC1 shows that the motifs detected for the individual clusters are in good agreement (Supplementary Tables S2–S4). We note that, as it contains more elements and had a higher content of photosynthesis-related genes and operons, the predictions for CHPC1 may be more reliable than those for CHPC2 for predicting motifs involved in photosynthesis gene regulation.

Combined CHPCs. Since, in general, increasing the sample size can be expected to lead to an increase in the

Fig. 2. Graphical representation of the consensus sequence derived from the predicted PpsR binding motifs described in Table 2. The relative sizes of the letters indicate their likelihood of occurring at a particular position.

statistical information content, we merged the sets of loci for the two CHPC clusters and then searched for motifs in the combined upstream sequence set. Search parameters for the combined clusters were the same as those used for CHPC1. Not surprisingly, transcription-factor binding motifs found for the combined clusters are very similar to the motifs found using the data for individual CHPC clusters (Supplementary Tables S2–S4). As we expect the predictions based on larger sample sizes to have better statistical relevance, we base our subsequent discussion mostly on the results for the combined clusters.

PpsR binding motif. Earlier studies suggest that PpsR binds to the nucleotide sequence TGTN₁₂ACA (Gomelsky *et al.*, 2000; Lee & Kaplan, 1992). One of the motifs found during our searches of the combined clusters, TGTCA-[A/G]NN[A/C][A/T][A/T/C]N[T/C]TGACA (Supplementary Table S2), is in agreement with this earlier finding, but is significantly more refined than the previously published sequence (TGTN₁₂ACA). We therefore assigned this motif as the new predicted PpsR consensus sequence (Supplementary Table S2 and Fig. 2).

Using the set of 2096 upstream region sequences, the MAST program (Methods) was applied to search within the genome for genes potentially regulated by PpsR. The search resulted in the detection of 11 genes whose upstream sequences contain the new PpsR DNA binding motif. These 11 predicted PpsR-targeted genes, together with their fold changes between expression levels at 10 W m⁻² light intensity (without DMSO) versus expression under aerobic conditions (30 % O₂), are listed in Table 2. Ten of the 11 predicted genes are known to be regulated by PpsR (Choudhary & Kaplan, 2000; Moskvina *et al.*, 2005; Zeng *et al.*, 2003), an observation that supports our approach. Strikingly, with the exceptions of *argD* and *bchC*, nine of

Table 2. Genome-scale prediction of PpsR targeted genes

Probe set	Gene name	Gene function	Site sequence	Pos.†	FC‡
Amino acid metabolism					
RSP2008	argD	Acetylornithine aminotransferase	TGTCATTCCTTCCTGACA	63	−1.5
Photosynthesis					
RSP0263*	bchC	Bacteriochlorophyll synthesis	TGTCCAATAAAGTTGACA	99	8.3
RSP0265*	crtE	Carotenoid biosynthesis	TGTAAGAAAAAGTTGACA	4	22.6
RSP0266*	crtD	Carotenoid biosynthesis	TGTCAACTTTTTCTTACA	55	2.3
RSP0271*	crtI	Carotenoid biosynthesis	TGTAAACCTGACTAGACA	47	3.8
RSP0272*	crtA	Carotenoid biosynthesis	TGTCTAGTCAGGTTTACA	94	17.6
RSP0281*	bchE	Magnesium-protoporphyrin IX monomethyl ester oxidative cyclase	TGTCAACTGAAATGGACA	17	14.8
RSP0284*	bchF	Bacteriochlorophyllide hydratase	TGTCAAAGAAAATTGACA	176	16.5
RSP0314*	pucB	Light-harvesting B800/850 protein	TGTAAGTCAGAATTGACA	32	15.6
			TGTCAGCCAACACTGACA	152	
			TGTCAGCGCAATGTGACA	127	
RSP1556*	puc2B	Second copy of pucB	TGTCATGCCATGCAGACA	161	24.7
			TGTCAGGGATTTCGATACA	136	
Regulatory protein					
RSP0283*	ppaA	Regulator for photosystem formation	TGTCAATTCTGACTTACA	279	14.8
			TGTCAATTTTCTTTGACA	135	

*Genes previously reported (Choudhary & Kaplan, 2000; Zeng *et al.*, 2003) to be potential regulons of PpsR.

†Pos., the distance from the last base of the sequence site to the putative translation start codon of the gene.

‡FC, fold change between the expression levels at 10 W m⁻² light intensity (without DMSO) versus aerobic conditions was calculated as the ratio between the higher expression value and the lower value. If expression under aerobic conditions is upregulated versus 10 W m⁻² conditions, the fold change is shown as a negative number.

Table 3. Potential FnrL-targeted operons and genes of the CHPC1 and CHPC2 clusters

An underscore between the gene names indicates that they belong to the same operon, and the genes are ordered according to their positions in the operon.

Operon	Operon description	Site sequence	Position
<i>RSP2395</i>	Cytochrome <i>c</i> peroxidase	TTGATCTAAGTCAA	65
<i>gene2234_RSP0820</i>	Cytochromes <i>b561</i>	TTGATGCGGATCAA	70
<i>RSP0166</i>	DnaK suppressor protein	TTGACCTGCATCAA	52
<i>feoA2_RSP1819_RSP1818_RSP1817</i>	Fe ²⁺ transport	TTGATCCGCGTCAA	89
<i>hemN</i>	Coproporphyrinogen III oxidase	TTGATCCTGCGCAA	36
<i>bchEJGP</i>	Bacteriochlorophyll biosynthesis	TTGACATGCATCAA	54
<i>pucBAC</i>	Light-harvesting complex II	TTGTCCCTTTTCAA	229
<i>RSP2247_RSP2248</i>	Translation elongation factor	TTGATTCAGGTCAA	145
<i>RSP0466_RSP0465</i>	Unknown function	TTGACCTACATCAA	59
<i>RSP3341</i>	Unknown function	TTGATTGCGATCAA	58

the genes are encoded in operons or by genes belonging to the two CHPC clusters. We therefore conclude that PpsR is not a major regulator outside the photosynthesis genes in *R. sphaeroides*.

FnrL binding motif. The DNA consensus sequence that binds FNR of *E. coli* has been established as TTGA-TN₄ATCAA, and by analogy FnrL of *R. sphaeroides* is hypothesized to bind to the same sequence (Zeilstra-Ryalls & Kaplan, 1998). From the analysis of the combined clusters, a similar motif, TTGAT[C/T][C/T]N[G/C]-[A/G]TCAA, was detected (Supplementary Table S3). Among the elements of the combined clusters, searches using the MAST program detected the presence of this putative FnrL recognition motif in the upstream sequences of the *pucBAC*, *bchEJGP* and *hemN* operons (Table 3). All of these genes or operons, from genetic studies, are known to be regulated by FnrL (Oh *et al.*, 2000; Zeilstra-Ryalls & Kaplan, 1998), validating our methodology. Within the combined clusters, the FnrL recognition motif was also found to exist in seven other loci upstream of genes or operons listed in Table 3.

A genome-wide search of the 2096 loci revealed 40 that were found to have the FnrL binding motif (Supplementary Table S5). These motifs have been aligned and are represented as a consensus diagram (Fig. 3). Ten of these 40 loci are upstream of genes that belong within the two CHPC clusters. Among these 40 genes, 12 (indicated in the table by *) have been previously reported to be regulated by FnrL in *R. sphaeroides* or by FNR in *E. coli* (Kammiller *et al.*, 1993; Oh *et al.*, 2000; Zeilstra-Ryalls & Kaplan, 1995, 1998; Zeilstra-Ryalls *et al.*, 1997). We note that for the majority of the genes predicted to be regulated by FnrL we lack the experimental knowledge to confirm the correctness of our prediction. We also note that the genome of *R. sphaeroides* is GC-rich. As the FnrL DNA binding consensus sequence is AT-rich, unlike the PrrA case detailed below, we expect our computational approach to have a high true prediction rate.

As FNR/FnrL is known to be involved in the regulation of a wide range of biological functions (Kang *et al.*, 2005), having a list of FnrL-regulated genes with different functionalities is not a major concern; it actually supports the validity of our approach.

PrrA binding motif. The PrrA DNA binding motifs that were detected when the two clusters were examined independently and the motifs detected when the two clusters were combined are compared in Table 4. Although the motifs found in different loci sets are similar, there are noticeable differences at some of the nucleotide positions between detected motifs. The PrrA recognition motif found for the combined clusters is a mixture of the motifs observed in the loci searches for the individual CHPCs (Table 4). The first eight nucleotide positions of the motif for the combined clusters, T[G/A/C]CGACA-[C/G], and the subsequent eight positions, [T/A][C/A]T-GTCG[C/A], show best matches with the motifs from CHPC2 and CHPC1, respectively.

Unassigned motifs. There may be other, currently unknown regulators involved in the regulation of the photosynthesis genes. Our motif search using MEME identified additional DNA motifs that may encode novel

Fig. 3. Graphical representation of the consensus sequence derived from the predicted FnrL binding motifs described in Supplementary Table S5. For a description of the relative heights of the letters, see the legend to Fig. 2.

Table 4. Comparison of potential PrrA binding consensus sequences obtained in our study with the consensus sequences reported in the literature

Cluster	Position																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
(A) Motifs detected by the MEME program for the two clusters individually or for the combined clusters																	
CHPC1		G	C	G T	G TC	C	C AT	C	T G	C	T	GT	C	C	G T	C	
CHPC2	TG	N	C	G	A C	C	G A	C AG	G	C G	T	G	C A	TA	C G	G	A C
CHPC1 + CHPC2	T	G AC	C	G	A	C	A	CG	T A	C A	T	G	T	C	G	C A	A
(B) Variable gap motif detected by the BioProspector program for the combined clusters																	
CHPC1 + CHPC2	CT	G C	C	G	G	C G	Variable lengths (0–10 bp)					G	TA	C	G A	C A	
(C) Consensus sequences reported in the literature																	
RegR <i>(B. japonica)*</i>	N	G	N	G	AG	C	AG	T	T	N	N	G	N	C	G	C	N
RegA <i>(R. capsulae)†</i>		G	C T	G	C G	C G	G A	N	N	TA A	T A	N	N	C	G A	C	
Multiple organisms‡	T C	G	C	G	AG	C	A G	Variable lengths (AT rich)				G	N	C	G	C	

*Consensus sequence based on the 11 critical binding sites for RegR of *B. japonicum* (Emmerich *et al.*, 2000).

†Consensus sequence for RegA of *R. capsulatus* (Swem *et al.*, 2001).

‡Common PrrA/RegA/RegR consensus sequence across multiple organisms (Laguri *et al.*, 2003).

photosynthesis-gene transcription-factor binding sites. These motifs and their locations with respect to the genes that they may regulate are included in Supplementary Tables S6 and S7.

Further analysis of the PrrA recognition motif

The predicted DNA binding sequence of PrrA found using the MEME program is highly degenerate (Table 4). To determine if our results depended on the motif search algorithm, we utilized another program, BioProspector (Liu *et al.*, 2001), to repeat the search for the PrrA motif. An advantage of the BioProspector program is that it allows for variable-width pattern searches in which the investigated motif can have the form block1–gap–block2, where block1 and block2

refer to the two recognition elements directly contacted by a regulator. Both blocks have fixed widths and the intervening gap can be of variable length. In the search for the PrrA motif, a 6-[0-10]-5 search parameter was used, i.e. the widths of block1 and block2 were 6 and 5 bp, respectively, and the intervening spacing (gap) had a range of 0–10 bp. One of the detected motifs (Supplementary Table S8), [C/T][G/C]CGG[C/G]-gap-G[T/A]C[G/A][C/A], is almost identical to the PrrA motif that was found using the MEME program (Table 4). We therefore assigned this motif as the putative PrrA DNA binding sequence with a variable width. The only notable disagreement between the MEME and BioProspector results is for the fifth position, where A dominates the MEME motif while the BioProspector result is dominated by G (Table 4). Thus, the fifth position

of the PrrA motif is inconclusive from our results and, as both programs seem to perform equally well, we predict that both A and G are probable.

To further probe the characteristics of the predicted PrrA consensus sequence, we have also compared the PrrA DNA binding motifs that were found in our analysis with the consensus sequences that have been predicted by other groups in earlier biochemical studies (Emmerich *et al.*, 2000; Laguri *et al.*, 2003; Swem *et al.*, 2001). As shown in Table 4, PrrA DNA binding motifs that were detected in our analysis for the combined cluster dataset are in good agreement with the predictions made by other groups using different approaches (Emmerich *et al.*, 2000; Laguri *et al.*, 2003; Swem *et al.*, 2001). The most significant difference between our predictions and those of earlier studies is that rather than being non-specific, our analysis specifies that position 13 in the motif is either T or A. Implications of this close agreement between our new results and these earlier published results will be discussed later.

In our motif search using the BioProspector program, we looked for variable gap motifs where the gap ranged between 0 and 10 bp. The detected motif [C/T][G/C]CGG[C/G]-gap-G[T/A]C[G/A][C/A] was observed at 170 different DNA sites in loci belonging to the CHPC clusters. These 170 putative PrrA DNA binding sites were distributed among upstream sequences of 51 out of the 57 operons belonging to the clusters. We note that results obtained using the MEME and the BioProspector programs are in good agreement (Table 4), and therefore this observation is unlikely to be an artifact of an individual algorithm.

Fig. 4 shows the percentage distribution of the widths of intervening spacers among the 170 predicted PrrA binding sites. The most probable gap width is 5 bp (17%), which coincides with the distance (from position 7 to 11) in Table 4. This lies between the two inverted repeats in the

Fig. 4. Distribution of the widths of intervening spacers among the 170 predicted PrrA binding sites.

fixed-gap motif detected by MEME for the combined clusters. Although a gap that varies between 0 and 10 bp is probably too variable to be real, we opted for a large gap range in our motif search to be inclusive in the searches. As shown in Fig. 4, predictions for PrrA DNA binding motifs with very small and very large gaps occur less frequently than motifs with a 5 bp gap. However, these very large and small gaps still exist at a statistically significant number of places.

To further examine the gap in the predicted PrrA DNA binding sequence, we divided the 170 predicted PrrA DNA binding sites into 11 groups according to their gap widths. Binding sites with identical gap widths were grouped together and statistically analysed to obtain their corresponding consensus sequence. The consensus sequences in the two-block regions for each spacer width were very similar to the overall consensus sequence derived from the 170 DNA binding sites (data not shown). This suggests that although the spacer width might have evolved to vary significantly, the two recognition elements of the binding sites have been well conserved.

BioProspector detected putative PrrA binding sites in 51 loci. For each of these loci, the putative PrrA binding sequence that showed the best match to the motif [C/T]-[G/C]CGG[C/G]-gap-G[T/A]C[G/A][C/A] (Supplementary Table S8) was selected and the gap width analysed. The distributions of the gap widths for these best matches are depicted in Fig. 5. Among the 51 best-matching sites, the 5 bp gap had the highest frequency; however, other statistically significant gap widths also occur. Among the 51 loci, 11 were predicted by BioProspector to have only one putative PrrA binding site. The statistical distribution

Fig. 5. Distribution of the widths of intervening spacers among the 51 best-matching PrrA binding sites that were present in 51 loci.

of gap widths of these 11 PrrA binding sites is reported in Supplementary Fig. S2. Again, no single gap width was dominant.

A mutant homologue of PrrA, called RegA*, is found in the closely related organism *Rhodobacter capsulatus*. This mutant protein possesses DNA binding activity that is independent of its phosphorylation status (Du & Bauer, 1999). In DNase-footprint experiments, the locations of one binding site in the *cycA* P2 promoter region (Karls *et al.*, 1999), four binding sites within the *cbb_I* promoter-operator region (Dubbs *et al.*, 2000), and six binding sites within the *cbb_{II}* promoter-operator region (Dubbs & Tabita, 2003) from *R. sphaeroides* were detected. We used the PrrA DNA binding motif [C/T][G/C]CGG[C/G]-gap-G[T/A]C[G/A][C/A] predicted using the BioProspector program to guide the sequence alignment of the 11 experimentally determined PrrA binding sites (Dubbs *et al.*, 2000; Dubbs & Tabita, 2003; Karls *et al.*, 1999) from *R. sphaeroides*. Table 5 shows how experimentally determined binding sites align with the consensus motif. Interestingly, with the exception of *cbb_I site2* and *cbb_{II} site4*, experimentally detected binding sites either contain the sequence GCGNC in their first blocks or contain GNCGC in their second blocks, but not both simultaneously. Thus, the presence of only one of the recognition elements, i.e. either GCGNC or GNCGC (Laguri *et al.*, 2003), might be sufficient for PrrA binding. This hypothesis has been reinforced by recent footprinting studies (X. Zeng & S. Kaplan, unpublished results). The presence of different combinations of recognition elements in one binding site might provide a mechanism of adjusting the PrrA–DNA interaction strength to allow for differential expression of its target genes. Based on this hypothesis, we scanned the 170

putative DNA binding sites to search for those that had the motif GCGNC in their first block and/or GNCGC in their second block. The resulting 64 DNA sites are listed in Table 6. Interestingly, of the selected 64 DNA binding sites, only eight contain both GCGNC and GNCGC (indicated in Table 6). We note that of the operons listed in Table 6, the *pucBAC*, *puhA*, *bchEJGP* and *pufBALMX* operons have been shown to be activated by PrrA under oxygen-limiting conditions (Eraso & Kaplan, 1994; Oh & Kaplan, 2001). The correct detection of genes and operons that have experimentally been shown to be regulated by PrrA further supports our prediction method. In addition to photosynthetic functions, the genes and operons listed in Table 6 are known or predicted to be involved in electron transfer, metal-ion transport, transcription and translation, among others, adding further evidence for the role of PrrA as a global regulator of gene expression in *R. sphaeroides*. This was further reinforced from the results obtained by searching the 2096 loci with the PrrA motif generated by BioProspector (Methods). The motif which showed GCGNC in its first block and/or GNCGC in its second block with a 3–7 bp gap was detected to be present in 1285 loci.

Of the three regulators that we investigate in this study, PrrA differs from PpsR and FnrL in one aspect that has important statistical implications: unlike PpsR and FnrL, the DNA binding motif for PrrA is dominated by G/C nucleotides. Since the genome of *R. sphaeroides* is GC-rich (69%), any statistical sequence analysis for the motifs containing many C/G sites will suffer from the correspondingly lower information content of the genome. For this reason, we might expect our false-positive prediction rate for the genes regulated by PrrA to be considerably higher than that for the PpsR and FnrL cases.

Table 5. Occurrence of the [C/T][G/C]CGG[C/G]-gap-G[T/A]C[G/A][C/A] motif within 11 experimentally determined PrrA binding sites*

Binding site†	Block1	Gap (bp)	Block2
<i>cbb_I site1</i>	CGCATC	2	GCCGC
<i>cbb_I site2</i>	CGAGGG	6	GCTGC
<i>cbb_I site3</i>	TGCGAC	5	GACCT
<i>cbb_I site4</i>	AGCCGC	0	GGCGC
<i>cbb_{II} site1</i>	TGCGGC	5	GTCTT
<i>cbb_{II} site2</i>	CGCGAC	1	AACAG
<i>cbb_{II} site3</i>	CGCGAC	0	ATGAT
<i>cbb_{II} site4</i>	TGCCGG	5	TCCGC
<i>cbb_{II} site5</i>	TGAAGG	0	GCCGC
<i>cbb_{II} site6</i>	TGCAGG	2	GTCGC
<i>cycAP2</i>	TGCGGC	5	GTCAT

*Experimental PrrA binding sites are taken from the literature (Dubbs *et al.*, 2000; Dubbs & Tabita, 2003; Karls *et al.*, 1999).

†Four PrrA binding sites within the *cbb_I* promoter-operator region (Dubbs *et al.*, 2000) and six binding sites within the *cbb_{II}* promoter-operator region (Dubbs & Tabita, 2003) were detected in DNase-footprint experiments.

To predict the relative importance of the interplay of PpsR, FnrL and PrrA on a genome-wide scale, we combined our binding-site data for the three regulators. Fig. 6 shows the predicted overlap in their regulatory roles. Of the 11 predicted PpsR regulons, eight were among the 1285 possible PrrA targets, whereas for the 40 genes likely to be regulated by FnrL, 32 were potential PrrA targets. Of the 2096 loci examined, two genes, namely *pucB* and *bchE*, are predicted, solely on the basis of the motif searches, to be regulated by all three regulators. Genetic approaches support this conclusion (Oh *et al.*, 2000).

To try and determine how well the predictions for PrrA binding match the *in vivo* state, we grew wild-type and PrrA2 (*prpA*[−] mutant) under anaerobic dark DMSO conditions, conditions that had not been used in the clustering. We then compared the expression patterns of the wild-type to those of the *prpA*[−] mutant and found that 850 genes showed a difference in expression of ≥ 1.5 fold.

We then determined how many of the 1285 genes predicted to be regulated by PrrA showed changes in their expression pattern when compared to wild-type. We found that 523 genes showed a change of expression pattern of ≥ 1.5 -fold

Table 6. DNA sequences selected from the 170 possible PrrA binding sites that contain either the sequence NGCGNC and/or the sequence GNCGC in their first and second recognition blocks, respectively

Operon	Operon description	Block1	Gap	Block2	Pos.*
Chaperone protein					
<i>RSP0166</i>	DnaK suppressor protein	TGCGGC	3	GTGCGC	154†
		TGCGGC	9	GTGGA	196
CO dehydrogenase					
<i>RSP2879_RSP2878_RSP2877_RSP2876‡</i>	Carbon monoxide oxidation	AGCGGC	5	GTCAA	103
DNA replication, recombination and repair					
<i>RSP3361</i>	Putative restriction endonuclease or methylase	TGCGGC	5	GACGC	135†
		CGCGGC	5	GTAC	192
Electron transport chain					
<i>gene2234_RSP0820‡</i>	Cytochrome <i>b562</i>	TCCGAC	0	GACGC	102
		CGCGGC	10	GTGAC	325
<i>RSP2022</i>	Cytochrome <i>b</i>	CGCGGC	1	GACGC	131†
<i>RSP2395</i>	Cytochrome <i>c</i> peroxidase	CGCGAC	2	GTGCGC	131†
<i>RSP2808_RSP2807‡</i>	Signal peptide protein (<i>RSP2808</i>); cytochrome <i>b</i> (<i>RSP2807</i>)	TGCGGC	1	GACGA	33
Hypothetical					
<i>RSP2085</i>	Conserved hypothetical protein	TGCGGC	6	GACCC	46
<i>RSP2087</i>	Conserved hypothetical protein	TGCGGC	5	GTGGC	43
<i>RSP3341</i>	Hypothetical protein	TGCGGC	4	GTGCC	126
<i>RSP3706</i>	Hypothetical protein	CGCGGC	9	GTGCC	100
		AGCGAC	10	GTCCA	246
Metabolism					
<i>RSP0160_RSP0159_RSP0158_RSP0157‡</i>	dTDP-glucose dehydratase (<i>RSP0160</i>); UDP-glucose 4-epimerase (<i>RSP0159</i>); glycosyl transferase (<i>RSP0158</i>); beta-mannanase (<i>RSP0157</i>)	TGCGGC	3	GTCCA	435
		TGCGGC	4	GAGCC	509
		TGCGAC	9	GACGC	587†
		TGCAGG	3	GTGCGC	737
<i>RSP0428</i>	Nitrogen fixation protein	CGCAGT	3	GTGCGC	146
		CGCGGC	4	GTCAA	241
<i>RSP0476</i>	L-Fucose phosphate aldolase	TGCGAC	0	GTGCGC	61†
		CCCGGC	8	GTGCGC	121
		ACTGAC	0	GACGC	324
<i>RSP2086</i>	Tetracenomycin polyketide synthesis hydroxylase	CGCGAG	1	GACGC	105
<i>RSP2985_RSP2984‡</i>	5-Aminolevulinic acid synthase isozyme (<i>RSP2985</i>); 5-aminolevulinic acid synthase (<i>RSP2984</i>)	CGCGAC	9	GTGGA	8
		CGCGGC	10	GTCAA	114
<i>RSP3496</i>	Zinc carboxypeptidase A metalloprotease	CGCGGC	1	GAGGC	5
Photosynthesis					
<i>pufBALMX</i>	Light-harvesting complex (<i>pufBA</i>); reaction centre (<i>pufLM</i>); electron transfer (<i>pufX</i>)	CGCGGT	7	GTGCGC	51
Q	Spectral complex assembly	CGCGAC	6	GAGAA	94
		TGCAAC	9	GACGC	274
		CGCGAG	8	GACGC	677
		TGCGAC	5	GTCAA	856
<i>crtIB</i>	Carotenoid biosynthesis	AGCGGG	9	GTGCGC	91
<i>bchIDO</i>	Bacteriochlorophyll biosynthesis	CCCGGC	4	GTGCGC	177
		TCTGGC	2	GACGC	441
		CGCGGC	5	GTGGC	750
<i>bchEJGP</i>	Bacteriochlorophyll biosynthesis	CGCGGG	7	GTGCGC	69
<i>bchFNBHLM</i>	Bacteriochlorophyll biosynthesis	TACGAC	5	GTGCGC	104
		AGCGGC	2	GACAC	239

Table 6. cont.

Operon	Operon description	Block1	Gap	Block2	Pos.*
<i>puhA</i>	Photosynthetic reaction centre	CGCAGC	6	GACGC	729
		TGCGGC	4	GTCGA	756
		CATGGT	10	GTCGC	791
		CATGGG	6	GTCGC	888
		CGCGGC	8	GTGGC	932
<i>pucBAC</i>	Light-harvesting complex II	GGCGGC	6	GTCGC	184†
		TGCGGC	6	GTCCC	236
		TGCGGC	8	GTGCC	924
<i>RSP1556_RSP1557‡</i>	The second <i>pucBA</i> operon, i.e. <i>pucBA2</i>	TCCGGC	5	GTCGC	194
Regulatory protein <i>tspO</i>	Outer-membrane protein	CGCGGC	6	GAGGC	114
		TGCGGC	4	GAGCC	177
		CGCGGT	0	GACGC	448
		CGCGAC	6	GAGGA	516
		CGTGGG	4	GACGC	558
<i>ppaA</i>	Regulatory protein	CCCGGG	5	GTCGC	83
		TGCGGC	2	GAGCA	261
<i>RSP0752</i>	Putative transcriptional regulatory protein	CGTAGC	8	GTCGC	69
Translation <i>RSP2247_RSP2248‡</i>	Translation elongation factor	TCCGGC	3	GACGC	68
<i>RSP2386</i>	Translation initiation factor	CGCGGG	3	GTCGC	86
Transporter <i>feoA2_feoA1_feoB_RSP1817‡</i>	Fe ²⁺ transport	CGTGGG	9	GTCGC	257
		CGCGGC	1	GACGC	15†
<i>RSP0777_RSP0776_RSP0775‡</i>	Lipoprotein transporter subunits (<i>RSP0777</i> , <i>RSP0776</i>); cytochrome <i>c</i> (<i>RSP0775</i>)				
<i>RSP1476</i>	Heavy metal transport ATPase	CGCGAC	8	GACGA	7
		CGCGAC	9	GTGGA	200
<i>RSP3160_RSP3159_RSP3158‡</i>	Putative membrane protein (<i>RSP3160</i>); ABC transporter subunits (<i>RSP3159</i> , <i>RSP3158</i>)	TGCGGC	2	GAGCC	34

*Distance from the last base of the second block to the putative translation start codon of the first gene in an operon.

†DNA site that fits the pattern NGCGNC-gap-GNCGC, i.e. the site has both recognition elements.

‡An underscore between the gene names indicates that they belong to the same operon, and the genes are ordered according to their positions in the operon.

(significant), and 520 genes showed a change of expression pattern of <1.5 -fold (considered insignificant in our selection scheme). The remaining genes were called absent by the analysis software, i.e. their level of expression was too low to be detected. This suggests that our methodology predicted 523 of the 850 genes that showed a significant change in expression, while falsely identifying 520 genes.

Of the 523 genes considered significant, 193 showed decreased expression in PRR2 (predicted to be PrrA acting as an activator) but, surprisingly, 329 genes showed increased expression (PrrA acting as a repressor). This result was a surprise, as PrrA is generally thought to act as an activator. These microarray results have been confirmed

Fig. 6. Overlap between predicted regulons of PpsR, FnrL and PrrA at a genome scale.

in part by randomly selecting seven genes and carrying out Northern blot analysis (J. M. Eraso & S. Kaplan, unpublished results).

DISCUSSION AND CONCLUSION

PrrA of *R. sphaeroides*, RegA of *R. capsulatus* and RegR of *Bradyrhizobium japonicum* are all response-regulatory proteins, and they share a 55 % overall amino acid sequence identity (Masuda *et al.*, 1999). More strikingly, the putative helix–turn–helix DNA binding domains of PrrA and its homologues exhibit a 100 % amino acid sequence identity (Masuda *et al.*, 1999). Thus, it can be expected that DNA consensus sequences that bind PrrA and its homologues may be very similar. A study by Emmerich *et al.* (2000) has shown that within the 17 bp minimal RegR binding site in *B. japonicum*, 11 nucleotides have been identified to be critical for regulator binding. In another study, Swem *et al.* (2001) proposed a 15 bp RegA consensus sequence for *R. capsulatus*. Based on the results of these two previous studies, Laguri *et al.* (2003) performed a multiple alignment on a set of experimentally determined RegA*/RegR DNA binding sites from *R. sphaeroides*, *R. capsulatus* and *B. japonicum* to elucidate the sequence [T/C]GCG[A/G]C[A/G]-gap-GNCGC as the common PrrA/RegA/RegR consensus sequence across multiple organisms. Table 4 compares the PrrA consensus sequence found in our analysis to the motifs that have been reported in the literature (Comolli *et al.*, 2002; Emmerich *et al.*, 2000; Laguri *et al.*, 2003; Swem *et al.*, 2001). In their multiple sequence alignment study, Laguri *et al.* (2003) allowed the width of the gap to be variable with a maximal value of 4 bp and determined the PrrA/RegA/RegR consensus sequence as [T/C]GCG[A/G]C[A/G]-gap-GNCGC. A comparison among RegR critical binding sites, the RegA consensus sequence and the common PrrA/RegA/RegR consensus sequence (Table 4) reveals the conserved sequence pattern GCG[A/G/C]C[A/G]-gap-GNCGC, which is not surprising, since PrrA, RegA and RegR share 100 % sequence identity in their helix–turn–helix DNA binding domains (Masuda *et al.*, 1999). This conserved sequence pattern closely matches the motifs that we have detected using the MEME and BioProspector programs for the combined cluster dataset and supports the validity of our approach.

Using a combination of structure determination and sequence analysis methods, Laguri *et al.* (2003) presented a convincing argument that PrrA binds to the DNA as a homodimer, and that the width of the gap between the DNA sequences GCGNC and GNCGC directly contacted by the two monomers can be variable. The variable gap in the DNA recognition motif is supported by our results. However, the *R. sphaeroides* genome has a 69 % G + C content and the consensus sequence of PrrA [C/T][G/C]CGG[C/G]-gap-G[T/A]C[G/A][C/A] is also dominated by G and C, which makes the search for putative PrrA binding sites less reliable. Thus, as discussed by Laguri *et al.* (2003), although the gap width of the PrrA binding motif may be

variable, one should be cautious about predicted binding sites with very large or small gap widths. Such sites should be treated as either false-positive detections or very-low-affinity binding sites until confirming biological data are collected.

In addition to determining a variable-gap PrrA consensus sequence, we also refined the putative PpsR consensus sequence, and confirmed the putative FnrL consensus sequence. The detected consensus sequences for PpsR, FnrL and PrrA were then used to predict their potential regulons on a genome-wide scale. Of 11 genes detected by MAST to be regulated by PpsR (Table 2), 10 have been reported earlier to be potentially regulated by PpsR (Choudhary & Kaplan, 2000; Moskvina *et al.*, 2005; Zeng *et al.*, 2003). These 10 genes are indicated in Table 2 and their encoded products are all related to the photosynthetic metabolic function; examples are bacteriochlorophyll biosynthesis (*bch* genes), carotenoid biosynthesis (*crt* genes), light-harvesting complex (*puc* genes), and the regulation of photosynthesis genes (*ppaA*). Not surprisingly, these 10 genes showed increased mRNA abundance under anaerobic conditions with 10 W m⁻² light intensity compared with aerobic growth conditions (Table 2). This result can be explained in part by the observation that under anaerobic conditions with light, AppA inhibits the repression activity of PpsR and thus derepresses PpsR-targeted genes (Braatsch *et al.*, 2002; Masuda & Bauer, 2002). Among these 10 genes, we found that the upstream sequences of four genes contain two predicted PpsR binding sites. Interestingly, for *pucB* and *puc2B*, the two predicted PpsR binding sites are separated by 7 bp, whereas for the two divergently transcribed genes, *bchF* and *ppaA*, the two sites are separated by 126 bp. The presence of two binding sites in the upstream sequence might be explained by the tetrameric structure of PpsR (Gomelsky *et al.*, 2000). However, binding of a PpsR tetramer to two sites that are separated by 126 bp probably requires either DNA looping or possibly the binding of two tetramers to the target sites. The latter possibility has been described in *Bradyrhizobium*, where, under oxidizing conditions, PpsR binds as an octamer (Jaubert *et al.*, 2004). Thus it would be interesting to see the effect of deleting one of the two predicted binding sites in *R. sphaeroides*.

Very recently, two divergently transcribed genes involved in the early steps of tetrapyrrole biosynthesis, *haem* (RSP0680) and *hemC* (RSP0679), were experimentally identified to be targeted by PpsR (Moskvina *et al.*, 2005). Their observed PpsR binding sites are all located within the coding regions for these two genes. As we chose the upstream regions of the genes to search for the regulator DNA binding sites (see Methods), these sites were not included in our set of upstream region sequences and therefore could not be detected with our approach. These, however, are not real false negatives; lack of their detection is purely due to the DNA-region selection criteria used. Compared with the 10 detected photosynthesis-related genes, *argD*, which encodes acetyl-ornithine aminotransferase, does not show an apparent link

with photosynthesis. The expression level of *argD* under aerobic conditions is similar to that of *puc2B*. However, under photosynthetic conditions at 10 W m^{-2} , *puc2B* mRNA abundance increases 25-fold, whereas *argD* mRNA abundance decreases (1.5-fold), which cannot be explained by the role of AppA in mediating the repression activity of PpsR under anaerobic photosynthetic conditions. Thus, we suggest that *argD* may be a false-positive detection, or that PpsR may under certain conditions function as an activator (Jaubert *et al.*, 2004). Based on our motif-searching results, we conclude that PpsR in *R. sphaeroides* is to a large extent specific for the regulation of photosynthesis-related genes (Moskvina *et al.*, 2005).

Compared to the genes regulated by PpsR, 40 genes that are predicted to be regulated by FnrL exhibit a much broader range of biological functions, including photosynthesis, signal transduction, electron transport, redox homeostasis and translation elongation (Supplementary Table S5). Due to their diverse functions, the expression patterns of these 40 genes under photosynthetic growth conditions, compared with aerobic conditions, vary considerably (Supplementary Table S5) (Kang *et al.*, 2005). For example, all five photosynthesis-related genes (*bchE*, *pucB*, *hemN*, *hemZ* and *hemA*) have a significant increase in mRNA abundance (6- to 16-fold), while the two *aa3* oxidase subunits (*coxI* and *coxII*) show decreases in mRNA abundance by ninefold, whereas the expression of *fnrL* itself shows little change. Thus, FnrL can exert its anaerobic regulation both positively and negatively.

When compared with the genome-wide predictions of regulation by PpsR and FnrL, a much larger number of genes (1285 of 2096 genes) have the potential to be influenced by PrrA. Noting that we lack the benchmark data to compute the possible false-positive detection percentages, we stress that a sizeable percentage of the predictions of putative regulation by PrrA may be false. If they are not false positives, that such a large number of genes are predicted to be candidates for PrrA regulation is probably influenced by the variable gap widths (which were allowed to vary between 3 and 7 bp) and the high G + C content of the *R. sphaeroides* genome. It also suggests that PrrA may require a much less stringent sequence pattern for binding compared with the binding requirements for PpsR and FnrL. Such a finding may also suggest that PrrA has the potential to act in a much more global fashion than the other two regulators, FnrL and PpsR. This suggestion has been reinforced by DNA microarray experiments, in which the gene expression of *prpA*⁻ mutant PRRA2 and wild-type cells grown under dark DMSO conditions was compared. In these experiments, the transcription of at least 850 genes (showing a transcription difference of > 1.5-fold) was found to be affected by the absence of PrrA (J. M. Eraso & S. Kaplan, unpublished results). Although the number of detected genes depends on the used fold ratio cutoff, the finding that 523 of these genes were captured by our predictions confirms the expected global regulatory role for PrrA and in part validates the

methodology described here. The finding that for ~60 % of these genes PrrA may act as a repressor turns on its head the conventional idea of the role of PrrA as an activator. These findings clearly suggest that it performs as both a repressor and activator, with a slight leaning in favour of repression.

It is interesting to note that in the microarray comparison between wild-type and PRRA2, only 523 genes were captured by prediction compared to the 850 genes found experimentally. It might be expected that because of possible false positives our method would capture more, not fewer, than the 850 genes found by microarray analysis. However, in our method, we scanned sequences upstream of operons. In an operon, by definition, there are always fewer upstream sequences than genes; for example, an operon of four genes will only have one upstream sequence. Therefore, in the highly unlikely event that our predictions were perfect, we would always underestimate the number of genes regulated by PrrA. This problem is compounded by the fact that binding sites can be buried within the coding regions of 'stand-alone' genes (Moskvina *et al.*, 2005), and operons will also be missed in our method, resulting in an underestimation of genes controlled by a regulator.

As with all prediction methods, the user should be aware of the possibilities for error. In the case of PrrA, overestimation of binding sites may occur as a result of genome G + C composition and a highly redundant PrrA binding sequence coupled with its own high G + C composition. Underestimation in this case can occur because the number of upstream sequences is always less than the number of operons and hence coding regions in the genome. In addition, our method suggests genes that may be directly controlled by regulator binding. It misses completely all genes where the effect of the regulator is indirect, i.e. the regulator is the first step or an intermediate in a longer regulatory pathway.

PrrA, FnrL and PpsR are three major transcription regulators that control the expression of photosynthesis genes of *R. sphaeroides* in response to environmental stimuli. By selecting two clusters enriched for photosynthesis genes from the microarray clustering results and analysing the loci belonging to these two clusters, we obtained PpsR and FnrL consensus sequences, as well as a variable gap motif that is predicted to be recognized by PrrA of *R. sphaeroides*. By applying this approach to other clusters derived from the microarray data, it should be feasible to determine the consensus sequences recognized by transcription factors involved in regulating other biological processes. One of the main aims of this study is to use computational methods to identify a small number of targets to be investigated in future experimental studies. As our results show, the ability to determine the DNA binding sequences of the regulators of interest and the ability to do a whole-genome-level search for putative regulatory targets are useful filtering tools to direct future experiments towards a limited number of genes. Such computational approaches are also useful in

putatively distinguishing the profile of the transcriptional regulators, i.e. whether they control a small or large number of genes. This work is being extended in two ways: one involves the expression patterns obtained from genes using the microarray analysis of *R. sphaeroides* PpsR, FnrL and PrrA mutants; the second involves the direct examination by biochemical and genetic techniques of genes identified in this study as being subject to regulation by each of the three regulators. Such studies are now under way.

ACKNOWLEDGEMENTS

We thank Heidi J. Sofia for useful discussions. The Pacific Northwest National Laboratory is a multiprogramme national laboratory operated by Battelle for the US Department of Energy under contract DE-AC06-76RL01830. This work was supported by the Advanced Modelling and Simulation of Biological Systems Program of the Office of Advanced Scientific Computing Research of the Office of Science, US Department of Energy. This work was also supported by the US Department of Energy as a subcontract to S.K. (DOE grant no. DE-FG02-01ER63232).

REFERENCES

- Bailey, T. L. & Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the International Conference on Intelligent Systems for Molecular Biology, ISMB 2*, 28–36.
- Bailey, T. L. & Gribskov, M. (1998). Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* **14**, 48–54.
- Braatsch, S., Gomelsky, M., Kuphal, S. & Klug, G. (2002). A single flavoprotein, AppA, integrates both redox and light signals in *Rhodobacter sphaeroides*. *Mol Microbiol* **45**, 827–836.
- Choudhary, M. & Kaplan, S. (2000). DNA sequence analysis of the photosynthesis region of *Rhodobacter sphaeroides* 2.4.1. *Nucleic Acids Res* **28**, 862–867.
- Comolli, J. C., Carl, A. J., Hall, C. & Donohue, T. (2002). Transcriptional activation of the *Rhodobacter sphaeroides* cytochrome *c₂* gene P2 promoter by the response regulator PrrA. *J Bacteriol* **184**, 390–399.
- Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. (2004). WEBLOGO: a sequence logo generator. *Genome Research* **14**, 1188–1190.
- Du, S. & Bauer, C. E. (1999). DNA binding characteristics of RegA. A constitutively active anaerobic activator of photosynthesis gene expression in *Rhodobacter capsulatus*. *J Biol Chem* **274**, 16343–16348.
- Dubbs, J. M. & Tabita, F. R. (2003). Interactions of the *cbb_H* promoter-operator region with CbbR and RegA (PrrA) regulators indicate distinct mechanisms to control expression of the two *cbb* operons of *Rhodobacter sphaeroides*. *J Biol Chem* **278**, 16443–16450.
- Dubbs, J. M., Bird, T. H., Bauer, C. E. & Tabita, F. R. (2000). Interaction of CbbR and RegA* transcription regulators with the *Rhodobacter sphaeroides cbb_H* promoter-operator region. *J Biol Chem* **275**, 19224–19230.
- Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95**, 14863–14868.
- Elsen, S., Swem, L. R., Swem, D. L. & Bauer, C. E. (2004). RegB/RegA, a highly conserved redox-responding global two-component regulatory system. *Microbiology and Molecular Biology Reviews* **68**, 263–279.
- Emmerich, R., Strehler, P., Hennecke, H. & Fischer, H. M. (2000). An imperfect inverted repeat is critical for DNA binding of the response regulator RegR of *Bradyrhizobium japonicum*. *Nucleic Acids Res* **28**, 4166–4171.
- Eraso, J. M. & Kaplan, S. (1994). *prpA*, a putative response regulator involved in oxygen regulation of photosynthesis gene expression in *Rhodobacter sphaeroides*. *J Bacteriol* **176**, 32–43.
- Eraso, J. M. & Kaplan, S. (1997). Oxygen-insensitive synthesis of the photosynthetic membranes of *Rhodobacter sphaeroides*: a mutant histidine kinase. *J Bacteriol* **177**, 2695–2706.
- Gomelsky, M. & Kaplan, S. (1997). Molecular genetic analysis suggesting interactions between AppA and PpsR in regulation of photosynthesis gene expression in *Rhodobacter sphaeroides* 2.4.1. *J Bacteriol* **179**, 128–134.
- Gomelsky, M., Horne, I. M., Lee, H. J., Pemberton, J. M., McEwan, A. G. & Kaplan, S. (2000). Domain structure, oligomeric state, and mutational analysis of PpsR, the *Rhodobacter sphaeroides* repressor of photosystem gene expression. *J Bacteriol* **182**, 2253–2261.
- Jaubert, M., Zappa, S., Fardoux, J. & 7 other authors (2004). Light and redox control of photosynthesis gene expression in *Bradyrhizobium*. Dual roles of two PpsR*. *J Biol Chem* **279**, 44407–44416.
- Joshi, H. M. & Tabita, F. R. (1996). A global two component signal transduction system that integrates the control of photosynthesis, carbon dioxide assimilation, and nitrogen fixation. *Proc Natl Acad Sci U S A* **93**, 14515–14520.
- Kammler, M., Schon, C. & Hantke, K. (1993). Characterization of the ferrous iron uptake system of *Escherichia coli*. *J Bacteriol* **175**, 6212–6219.
- Kang, Y., Weber, K. D., Qiu, Y., Kiley, P. J. & Blattner, F. R. (2005). Genome-wide expression analysis indicates that FNR of *Escherichia coli* K-12 regulates a large number of genes of unknown function. *J Bacteriol* **187**, 1135–1160.
- Karls, R. K., Wolf, J. R. & Donohue, T. J. (1999). Activation of the *cycA* P2 promoter for the *Rhodobacter sphaeroides* cytochrome *c₂* gene by the photosynthesis response regulator. *Mol Microbiol* **34**, 822–835.
- Laguri, C., Phillips-Jones, M. K. & Williamson, M. P. (2003). Solution structure and DNA binding of the effector domain from the global regulator PrrA (RegA) from *Rhodobacter sphaeroides*: insights into DNA binding specificity. *Nucleic Acids Res* **31**, 6778–6787.
- Lee, J. K. & Kaplan, S. (1992). *cis*-acting regulatory elements involved in oxygen and light control of *puc* operon transcription in *Rhodobacter sphaeroides*. *J Bacteriol* **174**, 1158–1171.
- Li, C. & Hung Wong, W. (2001). Model-based analysis of oligo-nucleotide arrays: model validation, design issues and standard error application. *Genome Biology* **2**, 1–11.
- Li, C. & Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci U S A* **98**, 31–36.
- Liu, X., Brutlag, D. L. & Liu, J. S. (2001). BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. In *Pacific Symposium on Biocomputing*, pp. 127–138.
- Masuda, S. & Bauer, C. E. (2002). AppA is a blue light photo-receptor that antirepresses photosynthesis gene expression in *Rhodobacter sphaeroides*. *Cell* **110**, 613–623.
- Masuda, S., Matsumoto, Y., Nagashima, K. V., Shimada, K., Inoue, K., Bauer, C. E. & Matsuura, K. (1999). Structural and functional analyses of photosynthetic regulatory genes *regA* and *regB* from *Rhodovulum sulfidophilum*, *Roseobacter denitrificans*, and *Rhodobacter capsulatus*. *J Bacteriol* **181**, 4205–4215.
- Moskvin, O. V., Gomelsky, L. & Gomelsky, M. (2005). Transcriptome analysis of the *Rhodobacter sphaeroides* PpsR regulon: PpsR as a

master regulator of photosystem development. *J Bacteriol* **187**, 2148–2156.

Oh, J. I. & Kaplan, S. (1999). The *cbb₃* terminal oxidase of *Rhodobacter sphaeroides* 2.4.1: structural and functional implications for the regulation of spectral complex formation. *Biochemistry* **38**, 2688–2696.

Oh, J. I. & Kaplan, S. (2001). Generalized approach to the regulation and integration of gene expression. *Mol Microbiol* **39**, 1116–1123.

Oh, J. I. & Kaplan, S. (2002). Oxygen adaptation. The role of the CcoQ subunit of the *cbb₃* cytochrome *c* oxidase of *Rhodobacter sphaeroides* 2.4.1. *J Biol Chem* **277**, 16220–16228.

Oh, J. I., Eraso, J. M. & Kaplan, S. (2000). Interacting regulatory circuits involved in orderly control of photosynthesis gene expression in *Rhodobacter sphaeroides* 2.4.1. *J Bacteriol* **182**, 3081–3087.

Pappas, C. T., Sram, J., Moskvina, O. V. & 7 other authors (2004). Construction and validation of the *Rhodobacter sphaeroides* 2.4.1 DNA microarray: transcriptome flexibility at diverse growth modes. *J Bacteriol* **186**, 4748–4758.

Penfold, R. J. & Pemberton, J. M. (1994). Sequencing, chromosomal inactivation, and functional expression in *Escherichia coli* of *ppsR*, a gene which represses carotenoid and bacteriochlorophyll synthesis in *Rhodobacter sphaeroides*. *J Bacteriol* **176**, 2869–2876.

Qian, Y. & Tabita, F. R. (1996). A global signal transduction system regulates aerobic and anaerobic CO₂ fixation in *Rhodobacter sphaeroides*. *J Bacteriol* **178**, 12–18.

Roh, J. H. & Kaplan, S. (2002). Interdependent expression of the *ccoNOQP-rdxBHIS* loci in *Rhodobacter sphaeroides* 2.4.1. *J Bacteriol* **184**, 5330–5338.

Roh, J. H., Smith, W. E. & Kaplan, S. (2004). Effects of oxygen and light intensity on transcriptome expression in *Rhodobacter sphaeroides* 2.4.1. Redox active gene expression profile. *J Biol Chem* **279**, 9146–9155.

Sistrom, W. R. (1962). The kinetics of the synthesis of photopigments in *Rhodospseudomonas sphaeroides*. *J Gen Microbiol* **28**, 607–616.

Swem, L. R., Elsen, S., Bird, T. H., Swem, D. L., Koch, H. G., Myllykallio, H., Daldal, F. & Bauer, C. E. (2001). The RegB/RegA two-component regulatory system controls synthesis of photosynthesis and respiratory electron transfer components in *Rhodobacter capsulatus*. *J Mol Biol* **309**, 121–138.

Zeilstra-Ryalls, J. H. & Kaplan, S. (1995). Aerobic and anaerobic regulation in *Rhodobacter sphaeroides* 2.4.1: the role of the *fnrL* gene. *J Bacteriol* **177**, 6422–6431.

Zeilstra-Ryalls, J. H. & Kaplan, S. (1998). Role of the *fnrL* gene in photosystem gene expression and photosynthetic growth of *Rhodobacter sphaeroides* 2.4.1. *J Bacteriol* **180**, 1496–1503.

Zeilstra-Ryalls, J. H., Gabbert, K., Mouncey, N. J., Kaplan, S. & Kranz, R. G. (1997). Analysis of the *fnrL* gene and its function in *Rhodobacter capsulatus*. *J Bacteriol* **179**, 7264–7273.

Zeilstra-Ryalls, J., Gomelsky, M., Eraso, J. M., Yeliseev, A., O’Gara, J. & Kaplan, S. (1998). Control of photosystem formation in *Rhodobacter sphaeroides*. *J Bacteriol* **180**, 2801–2809.

Zeng, X., Choudhary, M. & Kaplan, S. (2003). A second and unusual *pucBA* operon of *Rhodobacter sphaeroides* 2.4.1: genetics and function of the encoded polypeptides. *J Bacteriol* **185**, 6171–6184.