2008

The cis-regulatory map of Shewanella genomes

Jiajian Liu
Washington University School of Medicine in St. Louis
Xing Xu
Washington University School of Medicine in St. Louis
Gary D. Stormo
Washington University School of Medicine in St. Louis

5376–5390 *Nucleic Acids Research, 2008, Vol. 36, No. 16* *Published online 13 August 2008*
doi:10.1093/nar/gkn515

The *cis*-regulatory map of *Shewanella* genomes

Jiajian Liu, Xing Xu and Gary D. Stormo*

Department of Genetics, Washington University School of Medicine, 660 S Euclid, Box 8232, St Louis, MO 63110, USA

Received May 27, 2008; Revised July 26, 2008; Accepted July 29, 2008

ABSTRACT

While hundreds of microbial genomes are sequenced, the challenge remains to define their *cis*-regulatory maps. Here, we present a comparative genomic analysis of the *cis*-regulatory map of *Shewanella oneidensis*, an important model organism for bioremediation because of its extraordinary abilities to use a wide variety of metals and organic molecules as electron acceptors in respiration. First, from the experimentally verified transcriptional regulatory networks of *Escherichia coli*, we inferred 24 DNA motifs that are conserved in *S. oneidensis*. We then applied a new comparative approach on five *Shewanella* genomes that allowed us to systematically identify 194 nonredundant palindromic DNA motifs and corresponding regulons in *S. oneidensis.* Sixty-four percent of the predicted motifs are conserved in at least three of the seven newly sequenced and distantly related *Shewanella* genomes. In total, we obtained 209 unique DNA motifs in *S. oneidensis* that cover 849 unique transcription units. Besides conservation in other genomes, 77 of these motifs are supported by at least one additional type of evidence, including matching to known transcription factor binding motifs and significant functional enrichment or expression coherence of the corresponding target genes. Using the same approach on a more focused gene set, 990 differentially expressed genes derived from published microarray data of *S. oneidensis* during exposure to metal ions, we identified 31 putative *cis*-regulatory motifs (16 with at least one type of additional supporting evidence) that are potentially involved in the process of metal reduction. The majority (18/31) of those motifs had been found in our whole-genome comparative approach, further demonstrating that such an approach is capable of uncovering a large fraction of the regulatory map of a genome even in the absence of experimental data. The integrated computational approach developed in this study provides a useful strategy to identify genome-wide *cis*-regulatory maps and a novel avenue to explore the regulatory pathways for particular biological processes in bacterial systems.

INTRODUCTION

Most of the genome-scale studies of bacterial regulatory systems have focused on the extensively studied model organism *Escherichia coli*. Despite the significant progress that has been made in identifying *cis*-regulatory elements in *E. coli*, 60–70% of the about 300 transcription factors (TFs) have not been characterized for their DNA-binding sites (1,2). Even less is known for the poorly studied microorganisms. As more bacterial genomes are being sequenced, identifying the *cis*-regulatory maps for these genomes remains challenging.

One common procedure for identifying putative *cis*-regulatory sites is to search for common motifs in the promoters of coregulated genes derived from microarray expression profiles or ChIP-chip experiments (3,4). A number of computational methods (5–8) using this strategy have been developed and successfully applied to identify *cis*-regulatory elements in both prokaryotic (9) and eukaryotic genomes (10,11). However, erroneously clustered 'co-regulated' genes often add noise to the step of searching for common motifs, and the limited amount of experimental data makes it difficult to extend these methods to the genome scale.

Over the past several years, several computational methods have been developed to identify DNA motifs on the genome scale in bacterial systems (12–16). Li *et al.* (12) presented an algorithm that extracts binding sites of regulatory proteins of a single genome without the knowledge of coregulated genes. This approach identified statistically significant dimer patterns and derived motif profiles from clusters of similar patterns. Using this method, they obtained about 160 distinct motifs from the *E. coli* genome. However, this approach had a low sensitivity: only one-third of the 60 characterized motifs of *E. coli* TF-binding sites (TFBS) were identified.

*To whom correspondence should be addressed. Tel: 314-747-5534; Email: stormo@genetics.wustl.edu

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Mwangi *et al.* (17) and Studholme *et al.* (18) performed similar analyses to predict DNA motifs in *Bacillus subtilis* and *Streptomyces coelicolor*, respectively. Using sequences from multiple genomes, the phylogenetic footprinting approach (19–21) provides an alternative way to identify regulatory elements of whole genomes without relying on experimental identification of coregulated genes. McCue *et al.* (22) applied whole-genome phylogenetic footprinting on *E. coli* and several γ-proteobacteria to find conserved motifs and employed a Bayesian clustering algorithm (13) to cluster motifs into distinct sets. Their study provided by far the most extensive collection of *cis*-regulatory elements in any bacterial genome. Similar strategies have also been applied on several other bacterial genomes (15,16,23,24) to find DNA regulatory motifs (see Supplementary Table 4). However, of the 181 motifs identified in *E. coli* using this method, 65% were shown to have only one or two target operons (http://www.people.fas.harvard.edu/~junliu/clust/). These results suggest that the coverage of the predicted *cis*-regulatory network is relatively low, considering that there are about 2700 transcription units (TUs) in *E. coli* (25). Several factors may contribute to affect the sensitivity and coverage of the phylogenetic approaches. First, species selection might not be optimal, as many *E. coli* DNA motifs may not be conserved in the other species. Second, algorithms for motif merging and clustering might be inefficient: conserved motifs initially identified by phylogenetic footprinting might be erroneously merged with other motifs in the clustering step, and some motifs in the final predictions might be redundant.

In this report, we present a new approach for genome-wide identification of *cis*-regulatory motifs in bacterial systems and its application in uncovering the *cis*-regulatory map of *Shewanella oneidensis*. We use two approaches to systematically identify DNA motifs and their target genes (regulons): one is to take advantage of the current abundant knowledge on the experimentally characterized transcriptional regulatory networks in *E. coli* to infer DNA motifs conserved in *S. oneidensis*; the other is to apply a new comparative genomics approach (Figure 1) that integrates phylogenetic footprinting, motif discovery with PhyloNet (21) and motif hierarchical clustering on multiple *Shewanella* genomes to predict novel DNA motifs on the genome scale. We assess our comparative approach by first examining whether the motifs identified using the five *Shewanella* genomes (*S. oneidensis*, *S. denitrificans*, *S. frigidimarina*, *S. amazonensis* and *S. baltica OS155*) are also conserved in the seven newly sequenced and distantly related *Shewanella* genomes (*S. sp W3-18-1, S. putrefaciens*, *S. loihica*, *S. woodyi ATCC51908, S. sediminis HAW EB3, S. halifaxensis HAW EB4* and *S. pealeana ATCC 700345*). Furthermore, we compare our predicted motifs with known motifs and analyze the functional enrichment and expression coherence (EC) of the target genes of the predicted motifs. We find that 64% of the predicted motifs are conserved in the distantly related *Shewanella* species and 77 of these motifs are supported by at least one type of evidence. Applying our comparative approach on a focused gene set, the differentially expressed genes derived from microarray expression profiles of *S. oneidensis* during exposure to various metal ions, we were able to more sensitively identify regulatory motifs and their target genes potentially involved in metal reduction processes. Moreover, we find that our whole-genome comparative analysis is able to discover most of these motifs and their target genes, although in the absence of any experimental data.

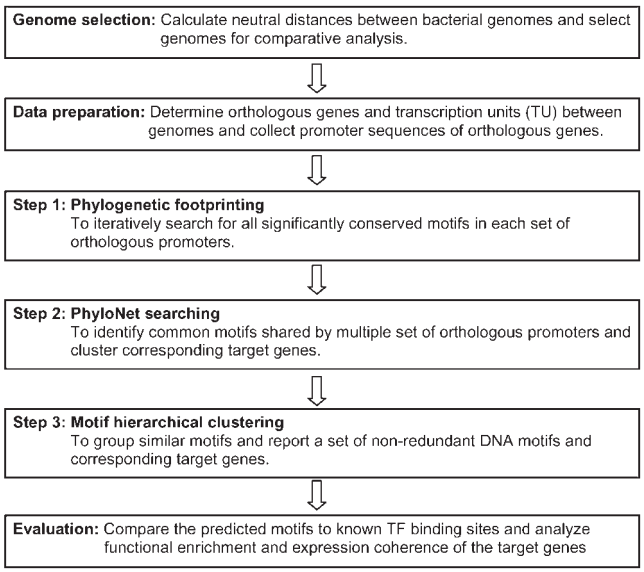


Figure 1. Flow chart of the procedure of identifying conserved *cis*-regulatory motifs in *S. oneidensis* by comparative analysis.

Shewanella oneidensis is a facultative, gram-negative γ-proteobacterium that can live in a wide variety of environments (26). Under anaerobic conditions, *S. oneidensis* can reduce various compounds, such as oxidized metals, inorganic chemicals and organic molecules (26–28). Its diverse respiratory capabilities ensure the great potential of *S. oneidensis* in bioremediation of both metal and organic pollutants. Although, many experimental studies are ongoing to determine the biological and biochemical characteristics of *S. oneidensis*, this organism is still far from being well understood. The complete genomic sequences of multiple *Shewanella* species provide valuable resources for DNA motif and regulon discovery using comparative genomics approaches. Identifying the *cis*-regulatory map in *S. oneidensis* will accelerate our understanding of the metabolism and gene regulation in this organism and ultimately facilitate its application in bioremediation.

MATERIALS AND METHODS

Datasets and genomic sequences

All complete genomic sequences used in this study were downloaded from the NCBI Genbank database. *Shewanella oneidensis MR-1*, *S. denitrificans OS217*, *S. frigidimarina NCIMB 400*, *S. amazonensis SB2B*, *S. baltica OS155* and *E. coli K12* were used for identifying *cis*-regulatory motifs in *S. oneidensis*. Eleven additional, recently sequenced *Shewanella* genomes (*Shewanella sp. ANA-3*, *S. baltica OS185*, *S. loihica PV-4*, *Shewanella*

sp. MR-4, *Shewanella sp. MR-7*, *S. putrefaciens CN-32*, *Shewanella sp. W3-18-1*, *S. woodyi ATCC51908*, *S. sediminis HAW EB3*, *S. halifaxensis HAW EB4* and *S. pealeana ATCC 700345*) were used to assess the conservation of the predicted motifs in these species. The datasets of experimentally characterized TF-target interactions in *E. coli* were downloaded from RegulonDB (version 5.0, released in 2006) (2). The series of publically available *S. oneidensis* microarray expression profiles under various conditions were collected from previous publications (28–30) and the M3D database (31) (see Supplementary Table 1 for downloading information).

Identification of orthologous genes between genomes

We identified orthologous genes by aligning all protein sequences from *S. oneidensis* (the anchor genome) to those from the other species using the NCBI BLASTP program (version 2.0) (6). Two genes were defined to be orthologous if all of the following three conditions are met: (i) their protein sequences are reciprocal best BLASTP hits between the two genomes; (ii) the BLASTP *E*-value is $<1.0 \times 10^{-10}$; and (iii) the BLASTP alignment covers $\geq$60% of the length of at least one sequence. Similar criteria have been used in a previous study (9).

Preparation of the promoter sequences of orthologous genes

Because genes in bacteria are organized in operons and most DNA motifs are located in the upstream regions of the first genes of the operons, we only retrieved the promoter sequence of each bacterial TU for motif search. In this study, a TU is defined as a set of consecutive genes in the same orientation whose intergenic distances do not exceed 40 nt (25). We collected the intergenic sequences up to 400 nt from upstream of the first gene of each TU in the anchor genome and its orthologs in the other genomes (not including any coding sequence if the length of the intergenic sequence is shorter than 400 nt). We refer to the promoter sequences of orthologous genes as 'orthologous promoters' in this article.

Inference of *S. oneidensis* TFBS from known *E. coli* regulatory interactions

We downloaded five datasets from RegulonDB (2) (version 5.0, 2006), including TF–target gene pairs, TFBS, promoters, gene products and alignment matrices, to compile the catalog of regulatory interactions for all experimentally characterized TFs in *E. coli*.

To use known regulons in *E. coli* to infer those in *S. oneidensis*, we first examined the conservation of *E. coli* TFs and their target genes (the first genes of regulated operons) in *S. oneidensis* by identifying orthologs between the two genomes. For each *E. coli* TF–target gene pair, we scanned the position specific weight matrix of the TF-binding motif along the promoter sequences of the orthologous target genes in the five *Shewanella* genomes (*S. oneidensis*, *S. denitrificans*, *S. frigidimarina*, *S. amazonensis* and *S. baltica OS155*) using the program Patser-v3b (32) to identify potential TFBS in these species. One standard deviation below the average of the Patser scores for all known *E. coli* binding sites of the TF was used as a cutoff. A binding site of an *E. coli* TF was considered conserved in *S. oneidensis* only if the promoters of the target gene orthologs in *S. oneidensis* and at least one of the other four *Shewanella* genomes had the binding sites whose Patser scores were not less than the cutoff.

Estimation of neutral distances between genomes

We estimated the neutral distances between all sixteen publicly available *Shewanella* species by examining the synonymous substitutions in coding sequences. The phylogenetic trees generated from the 16S rRNA sequences were used to guide the measurement of base substitutions.

There were 1670 genes conserved across all the sixteen *Shewanella* species. For each set of the orthologous genes, we first aligned their protein sequences using the program CLUSTALW (33), then converted the alignment of protein sequences into that of the DNA sequences, and finally calculated the synonymous substitution rate (K_s) using the program Codeml (34). Because the quality of multiple protein sequence alignments deteriorates as the sequence identity decreases (35), we only chose those alignments deemed reliable (protein sequence identity >75%) to estimate the K_s rates. Similar criteria have been used in previous studies (36,37).

Identification of novel DNA motifs on the genome scale in *S. oneidensis*

Identifying conserved DNA motifs in orthologous promoters by phylogenetic footprinting. As the majority of TFs in bacteria bind to DNA as dimers (38), we focused our investigation on the DNA motifs with palindromic patterns. We iteratively performed phylogenetic footprinting using the program CONSENSUS-v6c (32) to exhaustively identify all possible conserved palindromic motifs in each set of orthologous promoters. In each iteration, we searched for motifs with lengths of 17, 18, 22, 23, 27, 28, 32, 33, 37 and 38 nt. The identified significant sites ($E < 10^{-10}$) were hen masked and the modified sequences were used for the next round of motif discovery until no more significant motifs could be found. All motifs were represented as position weight matrix profiles.

Finding the common motifs shared by multiple genes using PhyloNet. We used the core algorithm of PhyloNet (21) to identify motifs that are shared by multiple sets of orthologous promoters. PhyloNet is a BLAST-like algorithm for motif searching and comparing. It queries each motif profile obtained from the phylogenetic footprinting step against all other motif profiles to search for common motifs and groups the corresponding target genes. For each query, PhyloNet reports up to 10 significant motif profiles ($P < 1 \times 10^{-10}$) and their target gene sets.

Clustering similar motifs to generate nonredundant motifs. Because of exhaustive querying, PhyloNet may identify redundant motif profiles and target gene clusters. To remove the redundancy, we used hierarchical clustering algorithms to cluster similar motif profiles and merge corresponding target genes. The similarity of two motifs

was measured by the average log likelihood ratio (ALLR) score of the global alignment of the two motifs (8) as defined below:

$$\mathrm{ALLR} = \frac{\sum_{b=\mathrm{A...T}} n^i_b \ln(f^i_b/p_b) + \sum_{b=\mathrm{A...T}} n^i_b \ln(f^j_b/p_b)}{\sum_{b=\mathrm{A...T}} (n^i_b + n^j_b)}$$

where f^i_b and f^j_b are base frequencies at aligned positions, i and j, respectively, from two motif profiles; n^i_b and n^j_b are observed base counts at aligned positions, i and j, respectively; and p_b is the background base frequency. In general, the higher the ALLR score, the more similar the two motifs.

We implemented three hierarchical clustering methods, including the single-linkage, complete-linkage and average-linkage agglomerative algorithms, to cluster the motifs. Initially, every motif profile output by PhyloNet is the only element in each cluster. The clusters with similar motif profile element(s) are gradually merged together. In the single-linkage algorithm, the similarity of two clusters was measured by the ALLR score of the nearest elements between the two clusters. In the complete-linkage algorithm, the similarity was based upon the ALLR score of the farthest elements between the two clusters. In the average-linkage algorithm, the similarity was defined as the mean of ALLR scores of all possible pairwise comparisons of the elements between the two clusters. Two-motif clusters were merged if all the following conditions were met: (i) ALLR > 7.5; (ii) the aligned portion of two nearest (single-linkage algorithm) or farthest (complete-linkage algorithm) motif profiles between the two clusters was longer than 75% of any one of the motifs; (iii) the two clusters had at least two common target genes. These merging criteria were unlikely to merge dissimilar motif profiles based on assessment on motif profiles of known TFBS. The motif clusters kept merging the next closest cluster until no cluster pair met the merging criteria. Finally, we chose the motif profile with the most target genes to represent the whole cluster of motif profiles for further analysis. All unique target genes from different motif profiles within a cluster were combined as the target genes of the final motif profile.

Conservation of predicted DNA motifs in seven newly sequenced, distantly related *Shewanella* genomes

One assessment of our predictions considers the conservation of the predicted DNA motifs in the distantly related *Shewanella* genomes that were not used for motif identification. We collected all 16 *Shewanella* species whose genomic sequences are available, including the initial five species used for motif finding and 11 newly sequenced species, and calculated their pairwise neutral distances using the method described above. A *Shewanella* species is considered distantly related to the initial five species if its neutral distance to any of the five species and other selected distantly related species is not less than the minimum neutral distance between any two of the five species. Seven of the 11 newly sequenced species (*S. sp W3-18-1, S. putrefaciens*, *S. loihica PV-4*, *S. woodyi ATCC51908, S. sediminis HAW EB3, S. halifaxensis HAW EB4* and *S. pealeana ATCC 700345*) meet such criteria and are used to examine motif conservation (see Supplementary Figure 2 and Supplementary Table 5). For each predicted motif, we scanned its position weight matrix along the promoters of all orthologous target genes in the seven distantly related *Shewanella* species using the program Patser-v3b (32) to identify the best matching sites. A motif is considered conserved in a distantly related *Shewanella* species only if the Patser score of its best matching site in this species is not less than that in *S. oneidensis*. As a control for spurious matches to low-specificity matrices, we scanned the same motif matrices in the upstream intergenic sequences of a set of randomly selected genes (excluding the orthologous target genes of the predicted motifs) from the seven *Shewanella* genomes.

Biological significance of predicted DNA motifs

Comparison of predicted DNA motifs to known TFBS. The binding motifs of a total of 30 known *E. coli* TFs that have orthologs in *S. oneidensis* were collected from RegulonDB (2). Focusing on palindromic motifs in this study, we compared our predicted motifs only to the 13 conserved *E. coli* TF-binding motifs with palindromic patterns. Motif similarities were measured by the ALLR statistics (8) as described above.

Functional enrichment analysis on the putative target genes of predicted motifs. As genes regulated by the same TF often have related functions in the cell, to find supporting evidence on our predictions, we examined the functional enrichment of target genes sharing the same motif in the anchor species (*S. oneidensis*) based on Gene Ontology (GO) (39) and KEGG pathway (KG) (40) annotations. The cumulative hyper-geometric distribution (11,41) was used to calculate the *P*-value ($P_{\mathrm{GO/KG}}$) of observing at least k genes having the same GO/KG functional annotation term among the n target genes of a motif profile.

$$P_{\mathrm{GO/KG}} = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

where M is the total number of genes with the particular GO/KG term and N is the total number of genes that have GO/KG annotations. Since bacterial genes are organized in TUs (operons), when calculating the GO/KG functional enrichment of the target genes, we considered all genes within the target TUs as long as they are annotated in the GO/KG databases.

Gene expression correlation analysis on the putative target genes of predicted motifs. Genes regulated by the same TF are expected to have correlated expression patterns in a series of microarray experiments. We used the EC score that was initially developed by Pilpel *et al*. (42) to measure the expression correlation of a set of target genes sharing the same motif. For each set of target genes, only the genes (K) that displayed varied expression across conditions were collected for study. The EC score of a set of K target genes is defined as p/P, where p is the number of

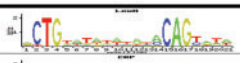
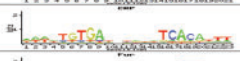
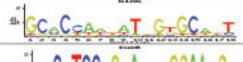
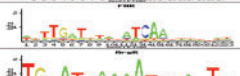
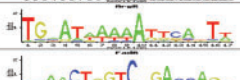
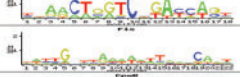
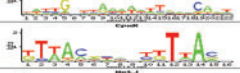
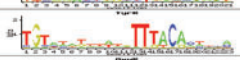
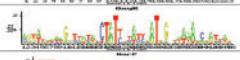
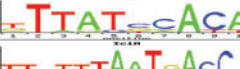
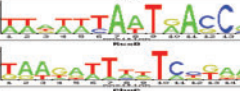
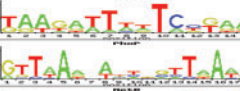
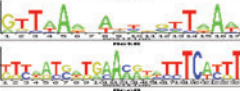
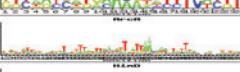
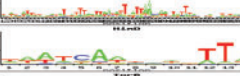
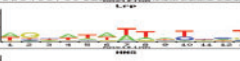
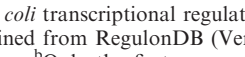
Index	*E. coli* TF	TF binding motif	Number of binding sites to build the profile [a]	Number of target genes in *E. coli* [b]	Number of *E. coli* target genes conserved in *S.oneidensis*	Number of conserved target genes retaining *E. coli* TF binding sites
1	LexA		23	13	12	5
2	CRP		182	118	38	23
3	Fur		47	25	5	1
4	GlnG		18	7	3	3
5	CueR		3	3	2	2
6	FNR		58	42	23	18
7	ArgR		17	6	3	1
8	FadR		10	7	5	1
9	Fis		117	24	18	14
10	CpxR		14	16	6	1
11	MetJ		23	7	6	4
12	PhoB		11	8	4	1
13	TyrR		17	8	5	0
14	OxyR		9	8	6	0
15	NarP		15	7	5	2
16	PspF		3	1	1	1
17	DnaA		8	5	4	2
18	IciA		4	2	2	1
19	RcsB		9	7	1	0
20	PhoP		17	15	10	0
21	MetR		4	3	3	3
22	ArcA		72	36	16	10
23	HimD		76	42	16	11
24	TorR		6	3	2	1
25	ModE		5	5	3	1
26	Lrp		54	16	5	3
27	HNS		29	3	1	1
28	OmpR		16	6	4	0
29	GcvA		4	2	2	2
30	CysB		8	5	4	0
			Total:	**450**	**215**	**112**

Figure 2. Conservation of the known *E. coli* transcriptional regulatory interactions in *S. oneidensis*. The logos of the *E. coli* TFBS were drawn based on the motif weight matrix models obtained from RegulonDB (Version 5.0). [a]The number of binding sites used to build the motif profiles. Multiple binding sites may be from the same gene. [b]Only the first genes of the target operons of a TF were considered when we counted the number of conserved target genes. A few target genes in RegulonDB were not included due to discrepancies in gene names.

gene pairs whose Euclidean distance between their mean and variance normalized expression profiles falls below a threshold, D (42); $P = 0.5 \times K \times (K-1)$, which is the total number of gene pairs in the set. We measured the significance of a resultant EC score by calculating the P-value (P_{EC}), the chance that a cluster of randomly selected K-qualified genes from the entire expression data has an equal or greater EC score than the real cluster of target genes in 1000 permutations. In this study, we collected expression data from three series of microarray experiments on *Shewanella*, including exposure of *Shewanella* to different electron acceptors (28,29), cold shock (30) and H_2O_2 shock (31) (see Supplementary Table 1). Each of these microarray studies had 8–12 conditions, and the expression profiles of target genes under all the conditions were used to calculate the EC scores. We considered that a cluster of target genes had significant EC if its P_{EC} <0.0183, which corresponds to a false discovery rate (FDR) of <0.05 (43), in at least one of these three microarray datasets.

Finding regulatory motifs and target genes involved in specific pathways from differentially expressed genes in microarray experiments

Differential gene expression in microarray experiments reflects cell's response to environmental changes or biological procedures. We can employ our comparative approach on orthologous promoters of differentially expressed genes derived from microarray experiments to identify regulatory motifs and their corresponding regulons that are potentially involved in specific biological processes. Instead of predicting *cis*-regulatory motifs individually from clustered genes that have similar expression patterns, our comparative approach is able to identify a set of putative regulatory motifs and corresponding target genes all at once. Moreover, working on a reduced motif search space, at a fixed significance level (the same P-value cutoff for PhyloNet to report motifs), this analysis should have a higher sensitivity in finding regulatory motifs than the whole genome analysis.

We applied this strategy on 990 differentially expressed genes derived from seven gene expression profiles of *S. oneidensis* during the exposure to a number of metal ions, including Fe(III), Mn(III), Mn(V), Cr(VI) and U(VI) (28,29), to identify regulatory motifs and corresponding target genes that are potentially involved in metal reduction.

RESULTS

Inferring regulatory interactions in *S. oneidensis* from verified transcriptional regulatory networks in *E. coli*

Shewanella oneidensis and *E. coli* both belong to the group of γ-proteobacteria and are assumed to share some common biology (26). We took advantage of the abundant knowledge on the *E. coli* genome and its regulatory networks and extrapolated it to the less characterized *S. oneidensis* genome.

We retrieved 62 *E. coli* TFs with known binding motifs and regulated genes (2) and examined the conservation of these TFs, their binding sites and target genes in *S. oneidensis*. We found that 30 of the 62 *E. coli* TFs have orthologs in *S. oneidensis*. Among the 450 TF–target pairs for these 30 TFs, 215 (48%) were conserved in *S. oneidensis* (Figure 2). We further examined whether the DNA-binding sites upstream of the *E. coli* target genes for the conserved TF–target pairs remain conserved in their counterparts in *S. oneidensis*. As shown in Figure 2, 24 of these 30 *E. coli* TFs had conserved DNA-binding sites in *S. oneidensis*. Of the 215 conserved TF–target gene pairs, 112 (52%) retained their TFBS in *S. oneidensis*. The list of 112 conserved regulatory sites and their positions in the *S. oneidensis* target genes are shown in Supplementary Table 2. These results suggest that ~25% of the overall *E. coli* experimentally characterized transcriptional regulatory interactions are well conserved in *S. oneidensis* with regard to TFs, target genes and TFBS. The knowledge about the regulation of these *E. coli* regulons can be useful for inferring regulatory interactions in the genome of *S. oneidensis*.

Sequence divergence of *Shewanella* species

The use of multiple *Shewanella* genomes (detailed information is shown in Supplementary Table 3) for phylogenetic footprinting allows a high coverage of orthologous genes and facilitates the discovery of DNA motifs specific to this clade of species. As the ability of comparative sequence analysis to uncover DNA sites is intimately tied to the neutral rate of the genome sequence evolution (44), we estimated the neutral divergence along the lineage of the five *Shewanella* species for motif discovery to represent their evolutionary distances. On aligned coding sequences of orthologous genes from any two *Shewanella* species (34), we measured the synonymous base substitution rates (K_s) in terms of nucleotide substitution events per synonymous site since their last common ancestor. Figure 3 shows the phylogenetic tree of the five *Shewanella* species based on their neutral distances. The branches of this tree are measured by the medians of the K_s rates on alignments of all orthologous gene sets. *Shewanella baltica OS155* is the closest species to *S. oneidensis*, with a divergence of about 0.65. This is almost the same as that between human and mouse, which have a divergence of about 0.64 (36,45,46). *Shewanella dentrificans* and *S. frigidiamarina* have similar distances of about 1.60 to *S. oneidensis*, more distant than that between *Caenorhabditis elegans* and *C. briggsae*, which have a divergence of 1.4 (36). *Shewanella amazonensis* is most distant to *S. oneidensis*, with a divergence of about 2.0, a distance slightly smaller than that between *Drosophila melanogaster* and *D. pseudoobscura*, which have a divergence of 2.4 (36). These species are sufficiently diverged that their orthologous intergenic regions are usually not globally well aligned, which means that conserved regulatory sites can be identified with low false positive and false negative rates using as few as three species (47).

Identifying novel DNA motifs in the genome of *S. oneidensis*

Using *S. oneidensis* as the anchor species, we generated two sets of sequence data for identifying conserved

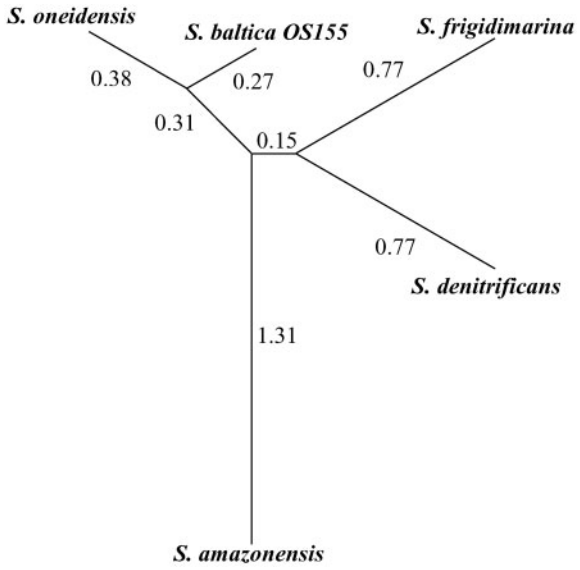


Figure 3. Estimation of neutral distances between five *Shewanella* species. The numbers shown are neutral substitution rates (K_s) measured by synonymous substitutions in coding sequences.

DNA motifs that are potentially *cis*-regulatory elements. Dataset I contains promoter sequences of orthologous genes from both *S. oneidensis* and *E. coli*, along with orthologs from at least one of the other four *Shewanella* species. Including sequences from *E. coli* added phylogenetic divergence which helped in identifying the most conserved *cis*-regulatory elements and also facilitated the functional annotation of our predictions on the poorly characterized *Shewanella* genomes. Dataset I has 862 sets of orthologous promoter sequences. Dataset II contains promoter sequences of orthologous genes from *S. oneidensis* and at least two other *Shewanella* species. It has 1961 sets of promoter sequences, of which 80% have at least four orthologous sequences. Because of the closer evolutionary distances between the *Shewanella* species, Dataset II covers more orthologous genes than Dataset I. By subtracting predicted motifs common to both datasets from those predicted from Dataset II, we can obtain *Shewanella* lineage specific regulatory interactions, which will help us understand how the regulatory networks in the *Shewanella* species make them special for metal reduction and anaerobic respiration.

Our comparative analysis for genome-wide motif discovery includes three steps as shown in Figure 1. Table 1 lists the numbers of input orthologous promoter sets and the numbers of motifs identified at each step. First, we performed phylogenetic footprinting iteratively using the program CONSENSUS-v6c (32) to exhaustively identify significantly conserved palindromic DNA motifs with different lengths in every set of orthologous promoter sequences. In this step, we found 11 247 and 22 525 conserved motifs from Dataset I and II, respectively, with an average of 11–13 motifs for each promoter set. This procedure is unlikely to miss any significantly conserved motifs, but it identifies many similar motifs (with different lengths) from the same orthologous promoter set. In a

Table 1. Numbers of motifs identified at each step of the comparative analysis for Dataset I and II

	Dataset I	Dataset II
The anchor genome(s)	*S. oneidensis* and *E. coli*	*S. oneidensis*
Other genomes	At least one other *Shewanella* species	At least two other *Shewanella* species
Number of sets of orthologous promoters	862	1961
Step I: Phylogenetic footprinting	11247	22525
Step II: PhyloNet searching	203	1665
Step III: Motif hierarchical clustering	38[a] (189)[b]	183 (824)

[a]Number of nonredundant motifs ultimately identified.
[b]Number of TUs covered by the predicted motifs.

regulatory network with a high degree of connectivity, each TF regulates multiple target genes, thus the *cis*-regulatory sites of the same TF can be found in multiple promoters. Hence, in the second step, we applied the core algorithm of PhyloNet (21) to identify common motifs shared by multiple promoters and cluster corresponding target genes. A large number of motifs that only occur in a single orthologous promoter set were discarded. After this step, the numbers of motifs in Dataset I and II were dramatically reduced to 203 and 1665, respectively (Table 1). However, the PhyloNet output motifs are still redundant, because PhyloNet is a BLAST-like algorithm for motif comparison and queries with different initial motif profiles may result in similar output motif profiles and target gene clusters. Therefore, in the third step, we further grouped similar motifs by hierarchical clustering methods. By assessing motif homogeneity within clusters and motif redundancy between clusters, we evaluated three agglomerative algorithms for motif clustering and found that the clustering results from the single-linkage algorithm were slightly better than those from the complete-linkage and average-linkage algorithms. Thus, we chose the motif profile clusters generated by the single-linkage algorithm as the final results in this article. In total, we identified 38 nonredundant motifs from Dataset I that covered 189 TUs, and 183 motifs from Dataset II that covered 824 TUs. The average numbers of target genes for the identified motifs were five and eight for Dataset I and II, respectively. Our predictions have a higher number of putative *cis*-regulatory interactions than those identified in previous studies in bacterial systems (12–18,22,23) (see Supplementary Table 4).

We obtained fewer motifs from Dataset I than from Dataset II, mainly because many *Shewanella* genes do not have orthologs in *E. coli* and some motifs are not well conserved between *E. coli* and *Shewanella* species. The fact that 38 motifs were conserved across *E. coli* and *Shewanella* species suggests that these motifs may play common and crucial roles in gene regulation in both genomes. Actually, the majority of motifs obtained from Dataset I were also identified from Dataset II. After eliminating the motifs common to both datasets, we

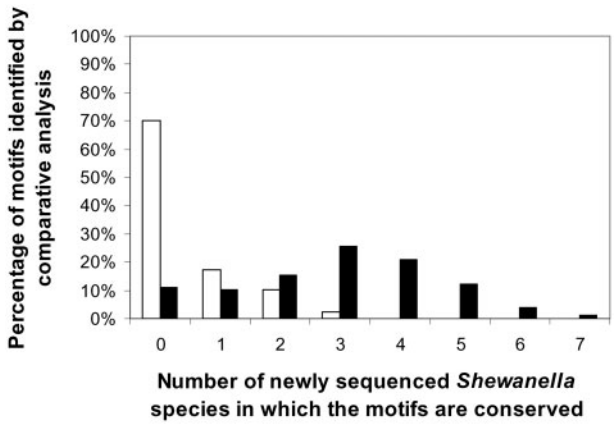


Figure 4. Conservation of the 183 motifs identified from Dataset II in the distantly related *Shewanella* species. The black bars represent the percentage of motifs identified from the five *Shewanella* genomes that are also conserved in the seven newly sequenced *Shewanella* species. The white bars represent the results from the control sequence sets.

obtained 155 motifs unique to Dataset II, which are likely to represent regulatory elements that are specific to the *Shewanella* lineage (Supplementary Table 6).

Evaluating the biological significance of predicted regulatory motifs

We first examined whether the predicted DNA motifs in *S. oneidensis* obtained through comparative analysis on the five genomes, including *S. oneidensis*, *S. denitrificans*, *S. frigidimarina*, *S. amazonensis* and *S. baltica OS155*, were also conserved in other newly sequenced *Shewanella* species. Of the 11 recently sequenced *Shewanella* species, seven (*S. sp W3-18-1, S. putrefaciens*, *S. loihica PV-4*, *S. woodyi ATCC51908, S. sediminis HAW EB3, S. halifaxensis HAW EB4* and *S. pealeana ATCC 700345*) were found distantly related to the initial five species used for motif discovery, based on their pairwise neutral distances and the criteria as described in Materials and methods section (Supplementary Figure 2 and Supplementary Table 5). These seven species were used to assess the conservation of the predicted motifs. Figure 4 shows that 64% of the 183 motifs identified from Dataset II were conserved in the orthologous promoters of at least three of the seven distantly related *Shewanella* species. In order to determine the background occurrence of the motifs, for each predicted motif, we checked its conservation in a control promoter set in which the orthologous promoters were replaced with the upstream intergenic sequences of the same number of randomly selected genes from the seven species. In only 2.2% of the control promoter sets was a motif conserved in at least three species. This suggests that the majority of our predicted motifs are evolutionarily constrained elements.

Because of the limited knowledge of TFBS in *Shewanella*, we chose to use the known *E. coli* TF-binding motifs to estimate the sensitivity of our predictions, which is defined as the fraction of known motifs that are correctly predicted. We used all 13 palindromic motifs from the 30 known binding motifs of *E. coli* TFs that have orthologs in *S. oneidensis* (2) to evaluate the sensitivity. Forty-six percent (6/13) and 69% (9/13) of the known motifs were correctly predicted from Dataset I and Dataset II, respectively (Supplementary Table 6). As shown in Figure 2, only a small fraction of the known *E. coli* transcriptional regulatory interactions were completely conserved in *S. oneidensis*. It is therefore not surprising that the prediction sensitivity for Dataset I was lower than that for Dataset II. Moreover, we compared the motifs identified from Dataset I with the predicted *E. coli* motifs that were computationally associated with TFs by Tan *et al.* (41). Four more motifs were supported by their study.

To further assess the biological significance of the predicted motifs, we analyzed the functional enrichment of their target genes (regulons) using the GO and KEGG pathway functional annotations and examined the EC of their target genes in various *Shewanella* microarray experiments. Of the total 183 motifs identified from Dataset II, the target genes of 24 motifs were significantly enriched for at least one biological function ($P_{GO/KG} < 1.0 \times 10^{-5}$) and the target genes of 46 motifs displayed correlated expression ($P_{EC} < 0.0183$ or FDR < 0.05) in at least one series of microarray experiments. Similarly, 6 and 13 of the 38 motifs from Dataset I were supported by the functional enrichment and EC of their target genes, respectively.

Figure 5 shows the numbers of the predicted motifs supported by the three types of evidence described above. Of the total 38 motifs identified from Dataset I, 17, 9 and 3 motifs were supported by at least one, two and three lines of evidence, respectively (Figure 5A and Table 2). Fifty-nine, 17 and 3 of the total 183 motifs identified from Dataset II were supported by at least one, two and three lines of evidence, respectively (Figure 5B and Table 3). The motifs with multiple lines of supporting evidence are the highest confidence predictions. Assuming that all the motifs with at least one type of supporting evidence are real motifs, we could estimate that the lower bounds of the specificities of our predictions are 45% (17/38) for Dataset I and 32% (59/183) for Dataset II, respectively. It is worth noting that even those motifs without any additional supporting evidence may be real, as 64% of them were conserved in at least three of the seven newly sequenced and distantly related *Shewanella* genomes.

By combining the 24 motifs inferred from known *E. coli* motifs, this study ultimately identified 209 unique *cis*-regulatory motifs in *S. oneidensis*. In addition to the motif conservation observed in the distantly related *Shewanella* genomes, 77 motifs (37%) were supported by at least one additional type of evidence. It should be pointed out that the numbers of motifs supported by EC and functional enrichment of target genes are underestimated, because many genes in *S. oneidensis* are unannotated and only a small number of expression datasets are available at this time.

Finding *cis*-regulatory motifs and target genes involved in metal reduction in *S. oneidensis*

Shewanella oneidensis is characterized by its diversified respiratory pathways, which ensure its capability of

using a wide variety of metal ion substrates as terminal electron acceptors in anaerobic respiration (26,29). We collected seven gene expression profiles of *S. oneidensis* under the conditions of exposure to a range of metal ions, including Fe(III), Mn(III), Mn(V), Cr(VI) and U(VI) (28,29), and identified 990 genes (in 460 TUs) that displayed at least 3-fold change in transcript abundance in at least one of the microarray conditions. From the orthologous promoters of these TUs in *S. oneidensis* and the other four *Shewanella* species, we identified 31 nonredundant DNA motifs and corresponding regulons. Four of these motifs matched the binding motifs of *E. coli* TFs: Fnr, Fur, FadR and MetJ. Three of these four TFs have been previously shown to be involved in anaerobic metabolisms or growth under iron limitation conditions in bacteria. Specifically, Fnr is known to control a large number of genes necessary for anaerobic metabolisms and the transition from aerobic to anaerobic respiration (27), and Fur regulates expression of genes encoding iron storage proteins and iron-utilizing enzymes under iron limitation (48). In addition, recent evidence indicates that FadR governs fatty acid metabolism through interaction with the TF, ArcA, during anaerobic growth (49). The target genes of six predicted motifs were enriched for at least one GO or KEGG functional term, the target genes of 12 motifs had correlated expression, and five motifs were supported by both lines of evidence. In total, 16 motifs were supported by at least one type of evidence and 71% of them were conserved in the newly sequenced and distantly related *Shewanella* species (Table 4).

We also identified several novel motifs that may be involved in regulating genes for metal reduction. First, the target genes of Motif 4 and Motif 5 shown in Table 4 were enriched for motor activity, flagellum and flagellar motility, and their expression profiles were highly correlated. Bacterial pili in both *Geobacter sulfurreducens* and *S. oneidensis* have been reported to serve as nanowires, transferring electrons from the cell surface to Fe (III) oxides (50–52). Whether the targets of these two motifs are involved in this process and how they function in metal reduction remain to be experimentally tested.

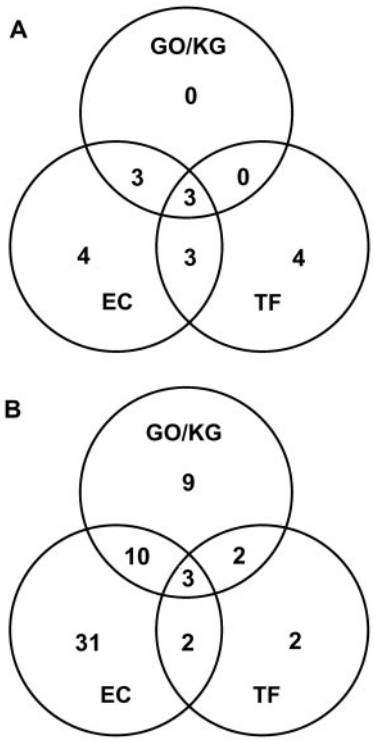


Figure 5. Venn diagrams showing the numbers of predicted motifs supported by different types of evidence, including matching to known TF-binding motifs, significant functional enrichment of the target genes in GO or KEGG pathway terms (GO/KG) and EC of the target genes in microarray experiments (EC). (A) Motifs identified from Dataset I (see Table 2 for detailed information). (B) Motifs identified from Dataset II (see Table 3 for detailed information).

Table 2. The list of predicted motifs from Dataset I that have at least one type of supporting evidence

Motif Number	Motif consensus sequence	Known TF	Expression coherence of target genes	Biological functions in GO(G) or KEGG pathway (K) terms for which the target genes are enriched
1	aCTGTwtaTAtawACAGt	LexA	Yes	DNA repair (G)
2	ACgTcTAgAcGTCtA	MetJ	Yes	Methionine biosynthesis (G)
3	tTGATctagATCAa	FNR	Yes	
4	yAaarNGCGCGCNyttTr		Yes	Structural constituent of ribosome (G), ribosome (G, K), protein biosynthesis (G)
5	ktAaAATkNcGCgNmATTtTam	CadC[a]	Yes	Protein biosynthesis (G), structural constituent of ribosome (G), ribosome (G, K)
6	AAtTtAAACgNNcGTTTaAaTT		Yes	Lipid catabolism (G, K)
7	TGTTGTaATATtACAACA		Yes	Pentose phosphate pathway (K)
8	aNTGaATtWWaATtCANt	ArgR		
9	tGGTcWgACCa	FadR		
10	tGCACcatwatgGTGCa	NtrC		
11	WAaaAWycgCGcgrWTttTW		Yes	
12	CACmAkATmTkGTG	YfeT[a]	Yes	
13	aatgCgScGcatt		Yes	
14	AtWTTgyatrcAAWaT		Yes	
15	GCGtAtWaTaCGC	Nlp[a]	Yes	
16	AAARggcgccwWwggcgccYTTT		Yes	
17	aTtGGTaWtACCaAt	PdhR[a]		

[a]Motifs that have been computationally associated with TFs by Tan *et al.* (41).

Table 3. The list of predicted motifs from Dataset II that have at least one type of supporting evidence

Motif Number	Motif consensus sequence	Known TF	Expression coherence of target genes	Biological functions in GO (G) or KEGG pathway (K) terms for which the target genes are enriched
1	ACTGTaTatawatACAGT	LexA	Yes	DNA repair (G)
2	TaGACGTCTAgA	MetJ	Yes	Methionine biosynthesis (G, K)
3	tTGATctagATCAa	FNR	Yes	Oxidoreductase activity, acting on NADH or NADPH, quinone or similar compound as acceptor (G), thiamine metabolism (K), other energy metabolism (K)
4	CCGtWaCGG	CueR		Acetolactate synthase activity (G), butanoate metabolism (K)
5	tGCACcawwwtgGTGCa	NtrC		Starch and sucrose metabolism (K)
6	aATGatAAtNaTTatCATt	Fur	Yes	
7	tGGTCWGACCa	FadR	Yes	
8	aGCYWRGCt		Yes	Structural constituent of ribosome (G), ribosome (G, K), rRNA binding (G), protein biosynthesis (G)
9	RGCANNwWwNNTGCY		Yes	Motor activity (G), ciliary or flagellar motility (G, K), flagellum (K, G)
10	tAaAATKNcGCgNMATTtTa		Yes	Protein biosynthesis (G), structural constituent of ribosome (G), ribosome (K), rRNA binding (G)
11	tCGCGa		Yes	Oxidoreductase activity, acting on NADH or NADPH, quinone or similar compound as acceptor (G), other energy metabolism (K)
12	GCGSCGC		Yes	Structural constituent of ribosome (G), ribosome (G, K), protein biosynthesis (G)
13	AgcGAcKRYMgTCgcT		Yes	Cytochrome complex assembly (G), heme transporter activity (G), nitrogen metabolism (K)
14	GTAATWWWATTAC		Yes	Glucose metabolism (G), main pathways of carbohydrate metabolism (G), pentose phosphate pathway (K)
15	tTtAAACaNNtGTTT		Yes	Lipid catabolism (G, K)
16	RGACAaWtTGTCY		Yes	Unlocalized protein complex (G), ferredoxin hydrogenase activity (G)
17	GTWATATWAC		Yes	Pentose phosphate pathway (K)
18	TGTAaaNNWWNNttTACA	TyrR		
19	TGaCANNatNNTGtCA	PhoB		
20	CGTrATyACG			Oxidoreductase activity, acting on NADH or NADPH, quinone or similar compound as acceptor (G), other energy metabolism (K)
21	CTSSAG			Hydrogen-transporting ATPase, ATP synthase activity, rotational mechanism (G, K), proton-transporting two-sector ATPase complex (G)
22	GCGyATWATrCGC			Oxidoreductase activity, acting on NADH or NADPH, quinone or similar compound as acceptor (G), other energy metabolism (K)
23	KaATATATtM			Cytochrome complex assembly (G), heme transporter activity (G)
24	GATcagGTTaA			Thiamin biosynthesis (G, K)
25	ccTGATCAgg			Thiamin biosynthesis (G, K)
26	CWCcSSgGWG			Bacterial chemotaxis (K)
27	GNGcMCAYKWTAWMRTGKgCNC			Cell division (K)
28	TGGCgaAtaTtcGCCA			Type II secretion system (K)
29	GACATAWTATGTC		Yes	
30	TATGSCATA		Yes	
30	TATGSCATA		Yes	
31	AcTTTACGTtaACGTAAAgT		Yes	
32	aAAAagSssscNwNgsssccttTTTt		Yes	
33	gGCsATsGCc		Yes	
34	aACNCSGNGTt		Yes	
35	CNatRMTSAKYatNG		Yes	
36	CCASTGG		Yes	
37	AAAGTSACTTT		Yes	
38	CMCCTWAGGKG		Yes	
39	tAccYgAgTaaWttAcTcRggTa		Yes	
40	ARAGYTARCTYT		Yes	
41	CAaTWAtTG		Yes	
42	RcATSATgY		Yes	
43	aAaAttkGcgCmaaTtTt		Yes	
44	CaAGGCCTtG		Yes	
45	AAATctGCwGCagATTT		Yes	
46	aacAtAAAGYNNRCTTTaTgtt		Yes	
47	TAAYRTTA		Yes	
48	TTtaTAcCTAGgTAtaAA		Yes	
49	gCCaAcAmaGSCtkTgTtGGc		Yes	
50	cgtWTGtTATAaCAWacg		Yes	
51	cTTCGAAg		Yes	

(continued)

Table 3. Continued

Motif Number	Motif consensus sequence	Known TF	Expression coherence of target genes	Biological functions in GO (G) or KEGG pathway (K) terms for which the target genes are enriched
52	TTTTRYAAAA		Yes	
53	WTGTSACAW		Yes	
54	GcMcTATATAgKgC		Yes	
55	gNcaaTGcTAgCAttgNc		Yes	
56	AaaAARCGYTTttT		Yes	
57	GGNAaAWTtTNCC		Yes	
58	AtgGCSGCcaT		Yes	
59	SCaTCawtGAtGS		Yes	

Table 4. The list of predicted motifs potentially involved in metal reduction that have at least one type of supporting evidence

Motif Number	Motif consensus sequence	Known TF	Expression coherence of target genes	Biological functions in GO (G) or KEGG pathway (K) terms for which the target genes are enriched
1	TaGACGTCTAgA	MetJ	Yes	
2	tGGTcWgACCa	FadR	Yes	
3	AGCcTAgGCT		Yes	Structural constituent of ribosome (G), ribosome (G, K), protein biosynthesis (G), rRNA binding (G)
4	RGCANNwWwNNTGCY		Yes	Flagellum, flagellar assembly (G, K), bacterial motility proteins (K)
5	cGTCaAaWWtTtGACg		Yes	Ciliary or flagellar motility (G), flagellum, flagellar assembly (G, K), bacterial motility proteins (K)
6	AgcGActRYagTCgcT		Yes	Heme-transporting ATPase activity (G), cytochrome complex assembly (G), nitrogen metabolism (K)
7	YAAATgaNAacsgtTNtcATTTR	Fur		
8	TGATctagATCA	FNR		
9	gggRggCTWAGccYccc			Structural constituent of ribosome (G), ribosome (G, K), protein biosynthesis (G), rRNA binding (G)
10	CcatTGSCAatgG		Yes	Copper ion binding (G), heme binding (G)
11	WAaNcGCGCgNtTW		Yes	
12	AaAGtTAaCTtT		Yes	
13	TGtAAcANyrNTgTTaCA		Yes	
14	aacAtAAAGYNNRCTTTaTgtt		Yes	
15	TTtAAACaNNtGTTTaAA		Yes	
16	aaAAarggNgcctNNgsssсctTTTt			Signal transduction mechanisms (K)

Second, the target genes of Motif 3 and Motif 9 were enriched for ribosome, protein biosynthesis and rRNA binding. The change of expression of genes in these functional categories during the transition from aerobic growth to the anaerobic condition has been observed previously (53). Third, the target genes for Motif 6 were enriched for heme-transporting ATPase activity and cytochrome complex assembly that involves ATP-binding cassette transporters. Genome analysis of *S. oneidensis* indicates that this organism encodes 43 putative c-type cytochrome genes (26,51). Two c-type cytochromes, MtrC and OmcA (51), have been implicated in electron transfer during metal reduction. The targets for Motif 6 provide new candidates for this process. Fourth, the target genes of Motif 10 were another set of c-type cytochrome genes that were involved in copper binding and heme-binding. Finally, Motif 16 had a large set of targets (14 TUs) that were enriched in a signal transduction pathway. Characterization of these genes might provide insights into how signal transduction is involved in metal reduction.

Eighteen of the 31 motifs discovered in this set of differentially expressed genes were also identified in the whole genome analysis. This indicates that the comparative analysis across multiple species at the whole genome level is capable of identifying a large fraction of the regulatory map of the genome in the absence of any experimental data. But using experimental information to identify subsets of genes involved in specific processes can increase the sensitivity of motif detection and allow us to further refine the regulatory pathways for particular biological processes.

DISCUSSION

By inferring conserved known TF-binding motifs from the well studied model microorganism *E. coli* and by systematically searching for novel DNA motifs on the genome scale, we identified 209 motifs that cover 849 unique TUs in *S. oneidensis*. Sixty-four percent of the predicted motifs are conserved in the seven newly sequenced, distantly related *Shewanella* genomes and 37% of them are supported by at least one of the three additional lines of evidence, including matching to known TF-binding motifs and significant functional enrichment or EC of

the target genes. In comparison with previous predictions (12–18,22,23) (see Supplementary Table 4), our study provides one of the most comprehensive *cis*-regulatory maps in bacterial systems with reasonable sensitivity and specificity. Moreover, by incorporating microarray profiling data into the analysis, our computational method also provides a novel strategy to characterize regulatory components for particular biological processes.

Inferring the regulatory networks from an extensively studied model organism is a direct way to identify conserved TFBS in a less characterized species. It also enables us to look into the evolution of the transcriptional regulatory networks across species. Several studies have been done on this topic (54,55). They are all based on the assumption that the regulatory interactions between a TF and its target genes in one species would occur in another species if both the TF and its target genes are conserved. This assumption might be true for those bacterial organisms that are evolutionarily close; however, it may not be appropriate for those evolutionarily distant genomes without taking into consideration the conservation of the TFBS. As shown in this study, for the 30 TFs that are conserved between *E. coli* and *S. oneidensis*, about 50% of the target genes in *E. coli* have orthologs in *S. oneidensis*, but only about 50% of these conserved targets bear the same TFBS. This observation implies that only about a quarter of the experimentally characterized transcriptional regulatory interactions in *E. coli* are still conserved in *S. oneidensis*. Hence, the previous estimate that about 40% of the *E. coli* transcriptional regulatory interactions are conserved in γ-proteobacteria (55) is likely an overestimate. Multiple factors can shape the bacterial transcriptional regulatory networks during evolution, such as gain or loss of TFs, target genes or TFBS, rearrangement of operon structures, and gene duplication through lateral gene transfer (56). Besides sharing the common targets, the orthologs of *E. coli* TFs in each *Shewanella* species may acquire new target genes to perform unique biological functions.

Species selection can significantly affect the effectiveness of comparative analysis in identifying DNA regulatory motifs in bacterial genomes. In order to cover more orthologous genes and discover regulatory elements specific to the *Shewanella* lineage, we chose five *Shewanella* genomes to identify DNA motifs genome wide. However, choosing evolutionarily closely related species can make it difficult to distinguish functional DNA-binding sites from background sequences, because the background sequences could be as conserved as the functional DNA sites. In this study, we used the genome-wide neutral mutation rates to estimate the evolutionary distances between bacterial species. Even though the five chosen *Shewanella* species are closely related across the bacterial phylogeny, they are sufficiently diverged so that three species are enough for identifying regulatory sites with low false positive and false negative rates (47,57). The result that the majority of the predicted motifs in *S. oneidensis* are also conserved in the seven newly sequenced, distantly related *Shewanella* genomes suggests that the predicted motifs evolved much slower than the background sequences. Therefore, it is likely that most of the predicted motifs are functionally constrained elements.

In addition to genome selection, several other factors can also significantly affect the quality of motif predictions. We applied an iterative search strategy for phylogenetic footprinting using the program CONSENSUS-v6c (32). This strategy allowed us to identify all possible significantly conserved motifs in each set of orthologous promoters and greatly enhanced the sensitivity of our predictions. However, many motif profiles identified from the phylogenetic footprinting step were highly redundant, and some of them might be false positives, appearing purely by chance in the orthologous promoters. Therefore, grouping similar motifs and eliminating false positives are essential to identifying true *cis*-regulatory motifs. As many false positive motifs are conserved in the orthologous promoters of a single gene and are not shared by other genes, PhyloNet (21) can efficiently eliminate a large number of false positive motifs by identifying common motifs shared by multiple promoters. With the subsequent step of motif merging using hierarchical clustering algorithms, we were able to further eliminate motif redundancy and generate a list of nonredundant motifs for the whole genome.

Our study provides a list of *cis*-regulatory motifs in *Shewanella* that are open for further validation. The detailed information of all predicted motifs is available online at http://ural.wustl.edu/databases.html. The fact that 64% of the predicted motifs are conserved in other distantly related *Shewanella* genomes and that at least 37% of the predicted *cis*-regulatory motifs are supported by additional evidence highlights the wealth of biologically significant information provided by our predictions. Furthermore, we have compared the motifs and target genes identified by the comparative genomics approach with those discovered from high-throughput microarray experiments in *S. oneidensis*. For example, from the microarray experiments on the *S. oneidensis* TF Fur and its mutant, Wan *et al.* (58) identified 21 coregulated target genes (operons) of Fur and predicted the consensus DNA motif of Fur binding sites from the promoters of these target genes using motif finding approaches. A motif profile identified by our comparative approach is very similar to that of the Fur binding motif obtained from the microarray data (Supplementary Figure 1A), and 11 of the 21 Fur target genes (52%) identified from microarray experiments were also shown in our predicted Fur regulon (Supplementary Figure 1B). Some of the Fur sites that we missed might be weak sites or specific to *S. oneidensis*, as they are not conserved in other species. Moreover, our approach predicted eight additional target genes that were not detected by the microarray experiments, and these genes are likely to be true targets, as they displayed strong EC with the 11 known target genes of Fur in a series of microarray experiments under stress conditions in which Fur is likely to be involved (59) (Supplementary Figure 1C). These results suggest that our comparative approach can extract useful biological information about regulatory networks in the absence of any experimental data, although having such data can provide a more informative view of the networks.

Although we expect that most of the predicted motifs are TFBS, we cannot rule out the possibility that some motifs may act as other functional elements. For example, some palindromic motifs could be part of the secondary structures of the RNA regulatory elements, such as riboswitches (60). However, considering the facts that the predicted motifs are all shared by multiple genes and that RNA regulatory elements usually are more conserved in secondary structure rather than in primary sequence (61), we think the number of RNA regulatory elements in our predicted motifs is very limited.

All predicted motifs and their target genes are putative components of the entire transcriptional regulatory networks. Using the nine motifs that match known TF-binding motifs, we can build the sub-networks containing TFs and their target genes. From such sub-networks, we can easily discover some useful biological information, such as the autoregulation of the TFs LexA, MetJ and NtrC, which has been reported in previous studies (62–64). With other sources of information, such as the distance constraint between TFs and their closest binding sites, phylogenetic correlation between TF and their regulons and binding specificity constraint for TFs (41), we may associate TFs with all predicted motifs to connect scattered components into a more complete transcriptional regulatory network that will allow us to gain insights into gene regulation in *S. oneidensis*.

We provide a strategy to identify bacterial lineage specific motifs by subtracting motifs conserved in more distantly related species (motifs from Dataset I) from all motifs identified in the lineage specific species (motifs from Dataset II). By this means, we predicted 155 *Shewanella*-lineage specific regulons, 43 of which are supported by additional evidence. Discovery of *Shewanella*-lineage specific regulatory interactions sheds light on the genomic differences that distinguish *Shewanella* from other γ-proteobacteria species and will help us understand how the distinct anaerobic respiration processes in *Shewanella* are regulated.

Instead of identifying *cis*-regulatory motifs individually from clusters of coregulated genes (11,42,65), our comparative genomics approach can predict a set of regulatory elements for particular biological processes all at once in the differentially expressed genes derived from microarray data. Our strategy can avoid problems that are often encountered in the former methods, such as limited expression conditions and incorrect or incomplete clustering of coregulated genes. With our approach, we successfully identified a set of known and novel *cis*-regulatory motifs involved in metal reduction in *S. oneidensis*. Those novel motifs can be high quality candidates for further experimental validation. Our predictions provide informative hints on which genes are possibly involved in the process of metal reduction and how they respond to environmental perturbations. These results show that utilizing experimental data can help elucidate not only regulatory motifs for specific biological processes, but also indicate that even in the absence of experimental data, comparative analysis using multiple related genomes can uncover a large fraction of the regulatory networks of an organism.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Dr Ting Wang for assistance with PhyloNet. We also thank Drs Timothy S Gardner and Rizlan Bencheikh-Latmani for providing the unpublished microarray data.

FUNDING

This work was supported by a grant from the US Department of energy (DE-FG02-05ER63972) to GDS.

Conflict of interest statement. None declared.

REFERENCES

1. Salgado,H., Gama-Castro,S., Martinez-Antonio,A., Diaz-Peredo,E., Sanchez-Solano,F., Peralta-Gil,M., Garcia-Alonso,D., Jimenez-Jacinto,V., Santos-Zavaleta,A., Bonavides-Martinez,C. *et al.* (2004) RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in Escherichia coli K-12. *Nucleic Acids Res.*, **32**, D303–D306.
2. Salgado,H., Gama-Castro,S., Peralta-Gil,M., Diaz-Peredo,E., Sanchez-Solano,F., Santos-Zavaleta,A., Martinez-Flores,I., Jimenez-Jacinto,V., Bonavides-Martinez,C., Segura-Salazar,J. *et al.* (2006) RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res.*, **34**, D394–D397.
3. Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
4. Johnson,W.E., Li,W., Meyer,C.A., Gottardo,R., Carroll,J.S., Brown,M. and Liu,X.S. (2006) Model-based analysis of tiling-arrays for ChIP-chip. *Proc. Natl Acad. Sci. USA*, **103**, 12457–12462.
5. Hertz,G.Z., Hartzell,G.W. III and Stormo,G.D. (1990) Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput. Appl. Biosci.*, **6**, 81–92.
6. Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.
7. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.
8. Wang,T. and Stormo,G.D. (2003) Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics*, **19**, 2369–2380.
9. Liu,J., Tan,K. and Stormo,G.D. (2003) Computational identification of the Spo0A-phosphate regulon that is essential for the cellular differentiation and development in Gram-positive spore-forming bacteria. *Nucleic Acids Res.*, **31**, 6891–6903.
10. MacIsaac,K.D., Wang,T., Gordon,D.B., Gifford,D.K., Stormo,G.D. and Fraenkel,E. (2006) An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics*, **7**, 113.
11. Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J. and Church,G.M. (1999) Systematic determination of genetic network architecture. *Nat. Genet.*, **22**, 281–285.
12. Li,H., Rhodius,V., Gross,C. and Siggia,E.D. (2002) Identification of the binding sites of regulatory proteins in bacterial genomes. *Proc. Natl Acad. Sci. USA*, **99**, 11772–11777.
13. Qin,Z.S., McCue,L.A., Thompson,W., Mayerhofer,L., Lawrence,C.E. and Liu,J.S. (2003) Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites. *Nat. Biotechnol.*, **21**, 435–439.
14. McCue,L.A., Thompson,W., Carmack,C.S. and Lawrence,C.E. (2002) Factors influencing the identification of transcription factor

binding sites by cross-species comparison. *Genome Res.*, **12**, 1523–1532.
15. Jensen,S.T., Shen,L. and Liu,J.S. (2005) Combining phylogenetic motif discovery and motif clustering to predict co-regulated genes. *Bioinformatics*, **21**, 3832–3839.
16. Alkema,W.B., Lenhard,B. and Wasserman,W.W. (2004) Regulog analysis: detection of conserved regulatory networks across bacteria: application to Staphylococcus aureus. *Genome Res.*, **14**, 1362–1373.
17. Mwangi,M.M. and Siggia,E.D. (2003) Genome wide identification of regulatory motifs in Bacillus subtilis. *BMC Bioinformatics*, **4**, 18.
18. Studholme,D.J., Bentley,S.D. and Kormanec,J. (2004) Bioinformatic identification of novel regulatory DNA sequence motifs in Streptomyces coelicolor. *BMC Microbiol.*, **4**, 14.
19. Blanchette,M. and Tompa,M. (2002) Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res.*, **12**, 739–748.
20. Siddharthan,R., Siggia,E.D. and van Nimwegen,E. (2005) PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput. Biol.*, **1**, e67.
21. Wang,T. and Stormo,G.D. (2005) Identifying the conserved network of cis-regulatory sites of a eukaryotic genome. *Proc. Natl Acad. Sci. USA*, **102**, 17400–17405.
22. McCue,L., Thompson,W., Carmack,C., Ryan,M.P., Liu,J.S., Derbyshire,V. and Lawrence,C.E. (2001) Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res.*, **29**, 774–782.
23. Wels,M., Francke,C., Kerkhoven,R., Kleerebezem,M. and Siezen,R.J. (2006) Predicting cis-acting elements of Lactobacillus plantarum by comparative genomics with different taxonomic subgroups. *Nucleic Acids Res.*, **34**, 1947–1958.
24. Conlan,S., Lawrence,C. and McCue,L.A. (2005) Rhodopseudomonas palustris regulons detected by cross-species analysis of alphaproteobacterial genomes. *Appl. Environ. Microbiol.*, **71**, 7442–7452.
25. Salgado,H., Moreno-Hagelsieb,G., Smith,T.F. and Collado-Vides,J. (2000) Operons in Escherichia coli: genomic analyses and predictions. *Proc. Natl Acad. Sci. USA*, **97**, 6652–6657.
26. Heidelberg,J.F., Paulsen,I.T., Nelson,K.E., Gaidos,E.J., Nelson,W.C., Read,T.D., Eisen,J.A., Seshadri,R., Ward,N., Methe,B. *et al.* (2002) Genome sequence of the dissimilatory metal ion-reducing bacterium Shewanella oneidensis. *Nat. Biotechnol.*, **20**, 1118–1123.
27. Beliaev,A.S., Thompson,D.K., Fields,M.W., Wu,L., Lies,D.P., Nealson,K.H. and Zhou,J. (2002) Microarray transcription profiling of a Shewanella oneidensis etrA mutant. *J. Bacteriol.*, **184**, 4612–4616.
28. Bencheikh-Latmani,R., Williams,S.M., Haucke,L., Criddle,C.S., Wu,L., Zhou,J. and Tebo,B.M. (2005) Global transcriptional profiling of Shewanella oneidensis MR-1 during Cr(VI) and U(VI) reduction. *Appl. Environ. Microbiol.*, **71**, 7453–7460.
29. Beliaev,A.S., Thompson,D.K., Khare,T., Lim,H., Brandt,C.C., Li,G., Murray,A.E., Heidelberg,J.F., Giometti,C.S., Yates,J. III *et al.* (2002) Gene and protein expression profiles of Shewanella oneidensis during anaerobic growth with different electron acceptors. *Omics*, **6**, 39–60.
30. Gao,H., Yang,Z.K., Wu,L., Thompson,D.K. and Zhou,J. (2006) Global transcriptome analysis of the cold shock response of Shewanella oneidensis MR-1 and mutational analysis of its classical cold shock proteins. *J. Bacteriol.*, **188**, 4560–4569.
31. Faith,J.J., Hayete,B., Thaden,J.T., Mogno,I., Wierzbowski,J., Cottarel,G., Kasif,S., Collins,J.J. and Gardner,T.S. (2007) Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol.*, **5**, e8.
32. Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.
33. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
34. Yang,Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.*, **13**, 555–556.
35. Katoh,K., Kuma,K., Toh,H. and Miyata,T. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.*, **33**, 511–518.
36. Pollard,D.A., Bergman,C.M., Stoye,J., Celniker,S.E. and Eisen,M.B. (2004) Benchmarking tools for the alignment of functional noncoding DNA. *BMC Bioinformatics*, **5**, 6.
37. Cooper,G.M., Brudno,M., Green,E.D., Batzoglou,S. and Sidow,A. (2003) Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. *Genome Res.*, **13**, 813–820.
38. Huffman,J.L. and Brennan,R.G. (2002) Prokaryotic transcription regulators: more than just the helix-turn-helix motif. *Curr. Opin. Struct. Biol.*, **12**, 98–106.
39. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.*, **25**, 25–29.
40. Ogata,H., Goto,S., Sato,K., Fujibuchi,W., Bono,H. and Kanehisa,M. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **27**, 29–34.
41. Tan,K., McCue,L.A. and Stormo,G.D. (2005) Making connections between novel transcription factors and their DNA motifs. *Genome Res.*, **15**, 312–320.
42. Pilpel,Y., Sudarsanam,P. and Church,G.M. (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.*, **29**, 153–159.
43. Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.
44. Cooper,G.M. and Sidow,A. (2003) Genomic regulatory regions: insights from comparative sequence analysis. *Curr. Opin. Genet. Dev.*, **13**, 604–610.
45. Cooper,G.M., Brudno,M., Stone,E.A., Dubchak,I., Batzoglou,S. and Sidow,A. (2004) Characterization of evolutionary rates and constraints in three mammalian genomes. *Genome Res.*, **14**, 539–548.
46. Nekrutenko,A., Makova,K.D. and Li,W.H. (2002) The K(A)/K(S) ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. *Genome Res.*, **12**, 198–202.
47. Eddy,S.R. (2005) A model of the statistical power of comparative genome sequence analysis. *PLoS Biol.*, **3**, e10.
48. Thompson,D.K., Beliaev,A.S., Giometti,C.S., Tollaksen,S.L., Khare,T., Lies,D.P., Nealson,K.H., Lim,H., Yates,J. III, Brandt,C.C. *et al.* (2002) Transcriptional and proteomic analysis of a ferric uptake regulator (fur) mutant of Shewanella oneidensis: possible involvement of fur in energy metabolism, transcriptional regulation, and oxidative stress. *Appl. Environ. Microbiol.*, **68**, 881–892.
49. Cho,B.K., Knight,E.M. and Palsson,B.O. (2006) Transcriptional regulation of the fad regulon genes of Escherichia coli by ArcA. *Microbiology*, **152**, 2207–2219.
50. Reguera,G., McCarthy,K.D., Mehta,T., Nicoll,J.S., Tuominen,M.T. and Lovley,D.R. (2005) Extracellular electron transfer via microbial nanowires. *Nature*, **435**, 1098–1101.
51. Marshall,M.J., Beliaev,A.S., Dohnalkova,A.C., Kennedy,D.W., Shi,L., Wang,Z., Boyanov,M.I., Lai,B., Kemner,K.M., McLean,J.S. *et al.* (2006) c-Type cytochrome-dependent formation of U(IV) nanoparticles by Shewanella oneidensis. *PLoS Biol.*, **4**, e268.
52. Kolker,E., Picone,A.F., Galperin,M.Y., Romine,M.F., Higdon,R., Makarova,K.S., Kolker,N., Anderson,G.A., Qiu,X., Auberry,K.J. *et al.* (2005) Global profiling of Shewanella oneidensis MR-1: expression of hypothetical genes and improved functional annotations. *Proc. Natl Acad. Sci. USA*, **102**, 2099–2104.
53. Rautio,J.J., Smit,B.A., Wiebe,M., Penttila,M. and Saloheimo,M. (2006) Transcriptional monitoring of steady state and effects of anaerobic phases in chemostat cultures of the filamentous fungus Trichoderma reesei. *BMC Genomics*, **7**, 247.
54. Madan Babu,M., Teichmann,S.A. and Aravind,L. (2006) Evolutionary dynamics of prokaryotic transcriptional regulatory networks. *J. Mol. Biol.*, **358**, 614–633.
55. Lozada-Chavez,I., Janga,S.C. and Collado-Vides,J. (2006) Bacterial regulatory networks are extremely flexible in evolution. *Nucleic Acids Res.*, **34**, 3434–3445.
56. Fraser-Liggett,C.M. (2005) Insights on biology and evolution from microbial genome sequencing. *Genome Res.*, **15**, 1603–1610.

57. Stone,E.A., Cooper,G.M. and Sidow,A. (2005) Trade-offs in detecting evolutionarily constrained sequence by comparative genomics. *Annu. Rev. Genomics Hum. Genet.*, **6**, 143–164.
58. Wan,X.F., Verberkmoes,N.C., McCue,L.A., Stanek,D., Connelly,H., Hauser,L.J., Wu,L., Liu,X., Yan,T., Leaphart,A. *et al.* (2004) Transcriptomic and proteomic characterization of the Fur modulon in the metal-reducing bacterium Shewanella oneidensis. *J. Bacteriol.*, **186**, 8385–8400.
59. McHugh,J.P., Rodriguez-Quinones,F., Abdul-Tehrani,H., Svistunenko,D.A., Poole,R.K., Cooper,C.E. and Andrews,S.C. (2003) Global iron-dependent gene regulation in Escherichia coli. A new mechanism for iron homeostasis. *J. Biol. Chem.*, **278**, 29478–29486.
60. Stormo,G.D. and Ji,Y. (2001) Do mRNAs act as direct sensors of small molecules to control their expression? *Proc. Natl Acad. Sci. USA*, **98**, 9465–9467.
61. Rivas,E. and Eddy,S.R. (2000) Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, **16**, 583–605.
62. Saint-Girons,I., Duchange,N., Cohen,G.N. and Zakin,M.M. (1984) Structure and autoregulation of the metJ regulatory gene in Escherichia coli. *J. Biol. Chem.*, **259**, 14282–14285.
63. Alvarez-Morales,A., Dixon,R. and Merrick,M. (1984) Positive and negative control of the glnA ntrBC regulon in Klebsiella pneumoniae. *EMBO J.*, **3**, 501–507.
64. Brent,R. (1982) Regulation and autoregulation by lexA protein. *Biochimie*, **64**, 565–569.
65. Spellman,P.T., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M.B., Brown,P.O., Botstein,D. and Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.