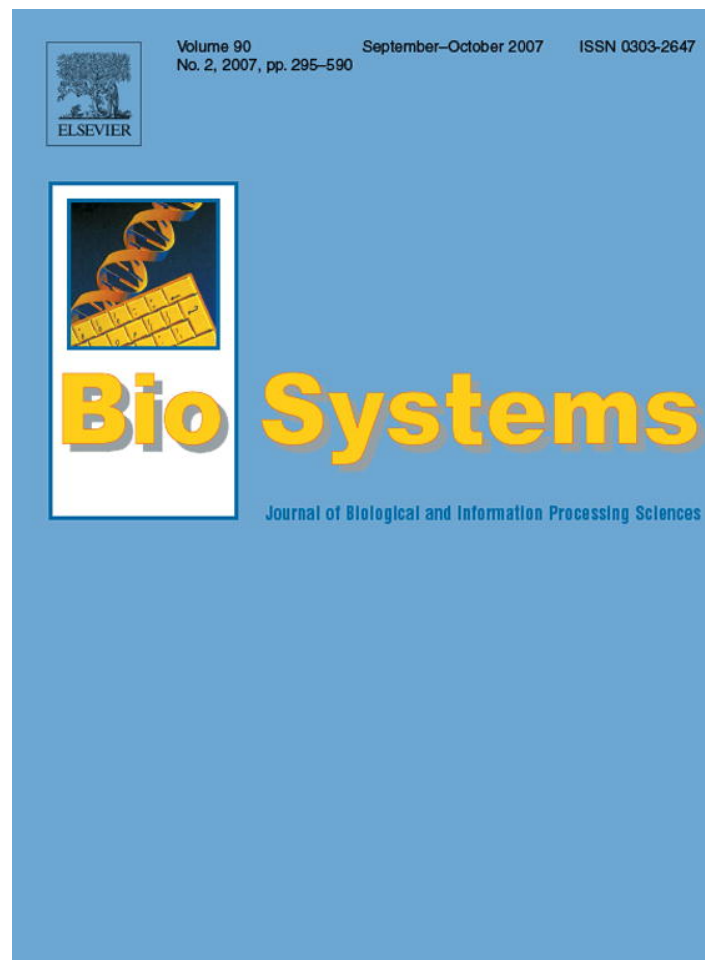


Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.



This article was published in an Elsevier journal. The attached copy is furnished to the author for non-commercial research and education use, including for instruction at the author's institution, sharing with colleagues and providing to institution administration.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/copyright>



ELSEVIER

Available online at www.sciencedirect.com

BioSystems 90 (2007) 421–441

www.elsevier.com/locate/biosystems

Genome-wide similarity search for transcription factors and their binding sites in a metal-reducing prokaryote *Geobacter sulfurreducens*

Bin Yan^{a,1}, Derek R. Lovley^b, Julia Krushkal^{a,*}^a Department of Preventive Medicine, University of Tennessee Health Science Center, 66 N. Pauline St., Ste. 633, Memphis, TN 38163, USA^b Department of Microbiology, Morrill Science Center IV North, University of Massachusetts, Amherst, MA 01003, USA

Received 27 July 2006; received in revised form 21 September 2006; accepted 20 October 2006

Abstract

The knowledge obtained from understanding individual elements involved in gene regulation is important for reconstructing gene regulatory networks, a key for understanding cellular behavior. To study gene regulatory interactions in a model microorganism, *Geobacter sulfurreducens*, which participates in metal reduction and energy harvesting, we investigated the presence of 59 known

Abbreviations: ABC, ATP-binding cassette; ATP, adenosine triphosphate; COG, clusters of orthologous groups; OmcB, outer membrane c-type cytochrome B; OmcC, outer membrane c-type cytochrome C; ORF, open reading frame; PID, GenBank protein identifier; S.D., standard deviation; Tu, transcription unit (singleton open reading frame not assigned to any operon); Ada, adaptive response to alkylating agents regulatory protein; AraC, arabinose operon regulatory protein; ArcA, aerobic respiration control protein ArcA; ArgR, arginine regulatory operon; CarP, aminopeptidase A/I, alternate name PepA; CpxR, transcriptional regulator CpxR; CRP, cyclic AMP receptor protein; CspA, cold shock protein 7.4; CynR, *cyn* operon positive regulator; CysB, positive transcriptional regulator for cysteine regulon; CytR, transcriptional regulator CytR; DeoR, transcriptional repressor for *deo* operon; DnaA, DNA-binding protein A; FadR, transcriptional regulator FadR; FarR, fatty acyl responsive regulator; FecI, RNA polymerase sigma factor (σ^{19}); FhlA, formate hydrogen-lyase transcriptional activator for *fdhF*, *hyc* and *hyp* operons; Fis, factor for inversion stimulation; FlhC, FlhC subunit of the FlhCD regulator of flagellar biosynthesis; FlhD, FlhD subunit of the FlhCD regulator of flagellar biosynthesis; FliA, RNA polymerase sigma factor (σ^F); Flp, FNR-like protein; FNR, FNR transcriptional dual regulator; FruR, FruR transcriptional dual regulator; Fur, ferric uptake regulator; GalR, repressor of *galETK* operon; GcvA, positive regulator of the *gcv* (glycine cleavage system) operon; GlnG, NtrC transcriptional dual regulator; GlpR, Repressor of the glycerol-3-phosphate regulon; HimA, integration host factor (IHF), alpha subunit; HimD, integration host factor (IHF), beta subunit; HipB, HipB transcriptional activator; H-NS, histoid-like nucleoid structuring protein; HU, heat-unstable nucleotide protein; HupA, DNA-binding protein HU, alpha subunit; HupB, DNA-binding protein HU, beta subunit; IclR, acetate operon repressor; IHF, integration host factor; IlvY, IlvY transcriptional dual regulator; LacI, transcriptional repressor of the *lac* operon; LexA, LexA transcriptional repressor; Lrp, leucine-responsive regulatory protein; MalT, positive regulator of mal regulon; MarR, repressor of *mar* operon; MelR, melibiose catabolism regulatory protein; MetJ, MetJ transcriptional repressor; MetR, MetR transcriptional activator; Mode, molybdate uptake regulatory protein; NagC, transcriptional repressor of *N*-acetylglucosamine operon; NarL, nitrate/nitrite response regulator NarL; NarP, nitrate/nitrite response regulator NarP; NtrC, NtrC transcriptional dual regulator, alternate name GlnG; OmpR, response regulator affecting transcription of *ompC* and *ompF*; outer membrane protein synthesis; OxyR, OxyR bifunctional regulatory protein sensor for oxidative stress; PepA, aminopeptidase A/I, alternate name CarP; PdhR, transcriptional regulator for pyruvate dehydrogenase complex; PhoB, unmodified phosphate regulon transcriptional regulatory protein PhoB; PurR, PurR transcriptional repressor; RhaS, positive regulator for *rhaBAD* operon; RpoD, RNA polymerase sigma factor (σ^{70}); RpoE, sigma factor (σ^{24}); RpoH, sigma factor (σ^{32}); RpoN, sigma factor (σ^{54}); RpoS, RNA polymerase sigma factor (σ^s , or σ^{38}); SoxS, regulator of superoxide response regulon; TorR, transcriptional dual regulator TorR; TrpR, transcriptional repressor TrpR; Tus, terminus utilization substance; TyrR, transcriptional regulator of aromatic amino acids metabolism

* Corresponding author. Tel.: +1 901 448 1361; fax: +1 901 448 7041.

E-mail address: jkrushka@utmem.edu (J. Krushkal).¹ Present address: NIDCD, National Institutes of Health, Building 10, Magnuson CC, Room 5D55, 10 Center Dr., Bethesda, MD 20892, USA.

Escherichia coli transcription factors and predicted transcription regulatory sites in its genome. The supplementary material, available at <http://www.geobacter.org/research/genomescan/>, provides the results of similarity comparisons that identified regulatory proteins of *G. sulfurreducens* and the genome locations of the predicted regulatory sites, including the list of putative regulatory elements in the upstream regions of every predicted operon and singleton open reading frame. Regulatory sequence elements, predicted using genome similarity searches to matrices of established transcription regulatory elements from *E. coli*, provide an initial insight into regulation of genes and operons in *G. sulfurreducens*. The predicted regulatory elements were predominantly located in the upstream regions of operons and singleton open reading frames. The validity of the predictions was examined using a permutation approach. Sequence similarity searches indicate that *E. coli* transcription factors ArgR, CytR, DeoR, FlhCD (both FlhC and FlhD subunits), FruR, GalR, GlpR, H-NS, LacI, MetJ, PurR, TrpR, and Tus are likely missing from *G. sulfurreducens*. Phylogenetic analysis suggests that one HU subunit is present in *G. sulfurreducens* as compared to two subunits in *E. coli*, while each of the two *E. coli* IHF subunits, HimA and HimD, have two homologs in *G. sulfurreducens*. The closest homolog of *E. coli* RpoE in *G. sulfurreducens* may be more similar to Fecl than to RpoE. These findings represent the first step in the understanding of the regulatory relationships in *G. sulfurreducens* on the genome scale.

© 2006 Elsevier Ireland Ltd. All rights reserved.

Keywords: *Geobacter sulfurreducens*; *Escherichia coli*; Operon; Transcription regulation; Transcription factor binding site

1. Introduction

Recent successes of genome sequencing projects have made it possible to compare the whole genome sequence information from different organisms in order to unveil their regulatory interactions. Multiple studies have attempted to identify transcription factor binding sites using genome-wide searches in either prokaryotic or eukaryotic species using information from known footprinted transcription regulatory elements within the same species or from other species (Chen et al., 1995; Thieffry et al., 1998; Tan et al., 2001; Studholme and Pau, 2003; Zheng et al., 2003; Guía et al., 2005). The use of known footprinted motifs from one species to search the genome sequence of another species has provided successful binding site predictions and new insights into regulatory mechanisms. The success of this approach has suggested some amount of conservation of certain transcription factor recognition sites among divergent species, including lineages of *Proteobacteria* (Tan et al., 2001; Studholme and Pau, 2003; Rodionov et al., 2004; Guía et al., 2005). A similarity search using matrices of footprinted regulatory elements has an advantage of allowing one to rapidly scan the entire genome to identify potential promoters and regulatory elements.

When interpreting the results of a cross-species search for transcription regulatory elements in bacteria, it is important to consider the location of the predicted elements relative to the start of the operons and open reading frames (ORFs). The noncoding regions upstream of operons usually contain promoters and are enriched with transcription factor binding sites (Thieffry et al., 1998; Rhodius and LaRossa, 2003). However, internal promoters and regulatory motifs also occur in intergenic regions (Homuth et al., 1997; Thieffry et al., 1998; LeBlanc et

al., 1999; Vaillancourt et al., 2002) and, infrequently, they are found inside open reading frames (Kormanec and Farkašovsky, 1993; Froehlich et al., 1994). Based on this consideration, the proportion of regulatory sites predicted in the upstream regions of the operons as compared to the sequence elements predicted in other genome locations may serve as an indicator of how likely the predicted sequence sites are to be true transcriptional regulatory elements.

Another step in eliminating false positive results among transcription regulatory elements predicted using interspecies comparisons is to verify whether transcription factors for which the binding sites have been found are indeed present in the organism of interest. While many recent studies have focused on the conservation of metabolic pathways across phylogenetic spectra, often less attention is paid to the conservation of the transcription regulators on the genome scale. We addressed this question by employing a genome-wide search of a dissimilatory metal-reducing and sulfur-reducing bacterial species, *Geobacter sulfurreducens* (Caccavo et al., 1994; Methé et al., 2003). We surveyed the *G. sulfurreducens* genome sequence for the presence of homologs of multiple transcription factors and of their binding sites that have previously been reported in *Escherichia coli* (Robison et al., 1998), a member of the γ -subdivision of *Proteobacteria* (Blattner et al., 1997).

G. sulfurreducens is a member of the δ subdivision of *Proteobacteria* (Lonergan et al., 1996). It was originally isolated from surface sediments of a hydrocarbon-contaminated drainage ditch (Caccavo et al., 1994). Because species of *Geobacter* can use high molecular-weight-organic compounds as electron donors and heavy metals as electron acceptors, these organisms are important for bioremediation of environments contaminated

with metal, metalloid, and organic waste compounds (Lovley, 1997; Holmes et al., 2002; Lovley, 2003). *G. sulfurreducens* also has an important ability to generate electricity by directly transferring electrons to electrodes, suggesting it as a possible source of energy (Bond and Lovley, 2003).

As part of the effort to understand the genetic potential and gene regulation of this important microorganism (Lovley, 2002), we investigated the presence of 59 *E. coli* transcription factors and their potential binding sites in *G. sulfurreducens*. We also investigated the distribution of putative transcription regulatory elements relative to predicted operon structure in the *G. sulfurreducens* genome. Our study compared the presence of transcription regulators, their predicted recognition sites, and operon predictions in *G. sulfurreducens*, providing a framework for a complex analysis of its transcription regulatory interactions on a genome scale.

2. Materials and methods

2.1. Identification of transcription factors in the *G. sulfurreducens* genome using similarity searches

The genome of *G. sulfurreducens* is 3814139 bp long (GenBank accession number AE017180), and it is located on a single circular chromosome (Methé et al., 2003). We used BLAST similarity searches to identify known transcription factors present in *G. sulfurreducens*. As a query, we used 59 *E. coli* K-12 transcription factors listed in the DPInteract database², for which matrices of aligned footprinted binding sites have previously been reported³ (Robison et al., 1998) (Table 1). Following the notation of Robison et al. (1998), in this report, we use the term matrix to describe a set of aligned regulatory sequence elements recognized by a particular transcription factor.

Protein sequences of 59 *E. coli* transcription factors were obtained from the National Center for Biotechnology Information⁴. The GenBank accession number for the *E. coli* K-12 genome is NC_000913. Information about *E. coli* transcription factors and their abbreviations was obtained from DPInteract,² EcoCyc,⁵ and BacTregulators⁶ databases (Robison et al., 1998; Karp et al., 2002; Martinez-Bueno et al., 2004), and from the online supplement⁷ to a study by McCue et al. (2002).

Protein sequences of *E. coli* transcription factors were used to query the entire set of protein data for *G. sulfurreducens* present in GenBank by program blastp using default search parameters. For heterodimer transcription factors FlhCD, IHF, and HU (Laine et al., 1980; Weglenska et al., 1996; Prüß et al., 2003), sequences of both subunits were used in separate BLAST searches (Table 1). To reduce the number of nonhomologous hits, sequences with blastp *E*-values of 10^{-1} or less, as well as those with *E*-values of 10^{-3} or less, were noted. To identify the number of potential homologs for each *E. coli* protein, only one blastp hit with the lowest *E*-value was considered for each *G. sulfurreducens* protein with a separate GI entry in GenBank. If a blastp search of *G. sulfurreducens* protein data returned no homologs of an *E. coli* transcription factor, we verified its absence by searching the entire nucleotide sequence of the *G. sulfurreducens* genome using program tblastn with default parameters.

In order to determine which of the identified blastp best hits in *G. sulfurreducens* were potential orthologs of *E. coli* proteins, we used the INPARANOID online tool for ortholog identification⁸ (Remm et al., 2001), to perform a reverse BLAST2 protein similarity search of *E. coli* proteins using the *G. sulfurreducens* proteins as a query. Those transcription factors that did not have a homolog in *G. sulfurreducens* were excluded from the reverse searches. We noted as potential orthologs those transcription regulators that were the best hits of one another in bi-directional similarity searches of *E. coli* and *G. sulfurreducens* proteins.

2.2. Phylogenetic analysis of HU, IHF, and RpoS protein sequences

Protein sequences of *E. coli* and *G. sulfurreducens* homologs of HU, IHF, and RpoE subunits were collected using the National Center for Biotechnology Information BLAST with Microbial Genomes server⁹. Sequences of *E. coli* RpoE and subunits of HU and their best blastp hit homologs in *G. sulfurreducens* (Table 1) were used as queries in blastp similarity searches using default software parameters to find other homologs in both species. Sequences were aligned using the popular Clustal X v. 1.81 sequence alignment software (Thompson et al., 1997) using default alignment parameters. Sequence alignments can be viewed at the online supplement for this report (<http://www.geobacter.org/research/genomescan/>). Positions in sequence alignments that contained insertions or deletions were excluded from phylogenetic tree inference. Phylogenetic trees were inferred with the commonly used MEGA software v. 3.1 (Kumar et al., 2004) using the neighbor-joining method, with Poisson correction of protein distances for multiple substitutions. The significance of clustering was evaluated by bootstrap with 500 replications. The trees were rooted at their midpoint (Graur and Li, 2000; Hall, 2004).

² <http://arep.med.harvard.edu/dpinteract/>.

³ http://arep.med.harvard.edu/ecoli_matrices/.

⁴ *E. coli* K12, complete genome at the National Center for Biotechnology Information. <http://www.ncbi.nlm.nih.gov/genomes/framik.cgi?db=genome&gi=115>.

⁵ <http://ecocyc.org/>.

⁶ <http://www.bactregulators.org/>.

⁷ http://bayesweb.wadsworth.org/binding_sites/study2.html.

⁸ <http://inparanoid.cgb.ki.se/>.

⁹ <http://www.ncbi.nlm.nih.gov/sutils/genom.table.cgi?database=83333>.

Table 1
Escherichia coli K-12 transcription factors and their homologs in the *Geobacter sulfurreducens* genome

Transcription factor	Location (bp) in <i>E. coli</i> genome	Strand	PID	Lowest blastp <i>E</i> -value in <i>G. sulfurreducens</i>	Accession number of the best blastp hit	No. of GI hits	No. of GI hits with $E \leq 0.1$	No. of GI hits with $E \leq 0.10^{-3}$
Ada	2307361–2308425	–	16130150	2×10^{-11}	NP_952226	5	1	1
AraC	70387–71265	+	16128058	1×10^{-6}	NP_954168	10	2	1
ArcA	4637159–4637875	–	16132218	2×10^{-31}	NP_954159	112	99	92
ArgR*	3382338–3382808	–	16131127	6.0	NP_954340	3	0	0
CarP	4482008–4483519	–	16132082	2×10^{-93}	NP_951392	8	1	1
CpxR	4102553–4103251	–	16131752	1×10^{-30}	NP_954159	107	94	81
CRP	3483757–3484389	+	16131236	3×10^{-13}	NP_954461	17	3	2
CspA	3717678–3717890	+	16131427	2×10^{-17}	NP_952954	9	4	4
CynR	357015–357914	–	16128323	3×10^{-40}	NP_953831	23	10	8
CysB	1331879–1332853	+	16129236	3×10^{-29}	NP_953831	17	8	7
CytR*	4121011–4122036	–	16131772	0.2	NP_951734	9	0	0
DeoR*	881199–881957	–	16128808	3.3	NP_952963	1	0	0
DnaA	3879954–3881357	–	16131570	6×10^{-17}	NP_952131	11	2	2
FadR	1234161–1234880	+	16129150	3×10^{-9}	NP_954436	6	4	4
FarR	764376–765098	–	16128705	1×10^{-22}	NP_951328	12	5	5
FhlA	2852361–2854439	+	16130638	1×10^{-76}	NP_952182	63	38	34
Fis	3408908–3409204	+	16131149	5×10^{-12}	NP_952057	28	15	4
FlhC*	1975290–1975868		16129843	0.2	NP_953563	10	0	0
FlhD*	1975871–1976230	–	49176166	3.5	NP_952912	5	0	0
FNR	1396798–1397550	–	16129295	4×10^{-14}	NP_954461	11	3	2
FruR*	88028–89032	+	16128073	0.33	NP_953970	9	0	0
Fur	709423–709869	–	16128659	10^{-23}	NP_952432	14	4	4
GalR*	2974621–2975652	+	16130741	0.58	NP_954440	9	0	0
GcvA	2939672–2940589	–	16130715	9×10^{-14}	NP_953831	11	8	8
GlpR*	3557480–3558238	–	16131297	0.23	NP_951328	9	0	0
HipB	1590200–1590466	–	16129467	0.004	NP_953586	11	4	0
H-NS*	1291732–1292145	–	16129198	0.95	NP_952447	3	0	0
HU (HupA)	4197859–4198131	+	16131830	2×10^{-25}	NP_954173	14	5	5
HU (HupB)	460675–460947	+	16128425	4×10^{-21}	NP_954173	8	5	5
IclR	4220383–4221246	–	16131844	7×10^{-31}	NP_951572	16	5	5
IHF (HimA)	1793277–1793576	–	16129668	2×10^{-23}	NP_952572	14	5	5
IHF (HimD)	963051–963335	+	16128879	2×10^{-19}	NP_953647	14	5	5
IlvY	3954548–3955441	–	16131631	2×10^{-16}	NP_953569	18	6	6
LacI**	365652–366743	–	33347444	0.02	NP_953747	14	1	0
LexA	4254694–4255302	+	16131869	3×10^{-34}	NP_952668	22	2	2
Lrp	931818–932312	+	16128856	0.03	NP_954410	7	1	0
MalT	3550718–3553423	+	16131294	3×10^{-5}	NP_954270	7	3	1
MarR	1617144–1617578	+	33347561	2×10^{-6}	NP_952534	12	3	3
MelR	4338298–4339206	–	16131944	6×10^{-5}	NP_954168	6	2	1
MetJ*	4125658–4125975	–	16131776	0.41	NP_953038	7	0	0
MetR	4009453–4010406	–	16131677	5×10^{-13}	NP_953569	13	7	5
ModE	793079–793867	–	16128729	2×10^{-23}	NP_954006	7	1	1
NagC	699597–700817	–	16128652	3×10^{-28}	NP_952753	6	1	1
NarL	1274402–1275052	–	16129184	4×10^{-33}	NP_952346	118	92	87
NarP	2288520–2289167	+	16130130	5×10^{-26}	NP_954270	105	86	79
NtrC	4051449–4052858	–	16131708	9×10^{-99}	NP_952057	123	110	102
OmpR	3533503–3534222	–	16131282	3×10^{-36}	NP_953194	109	99	94
OxyR	4156069–4156986	+	16131799	2×10^{-31}	NP_953831	14	7	7
PdhR	122092–122856	+	16128106	8×10^{-34}	NP_952677	23	6	5
PhoB	416366–417055	+	16128384	3×10^{-46}	NP_952155	106	98	93
PurR*	1735868–1736893	+	16129616	0.44	NP_952494	8	0	0
RhaS	4095317–4096153	+	16131745	4×10^{-9}	NP_954168	3	2	2
RpoD	3210688–3212529	+	16130963	1×10^{-137}	NP_954130	15	4	4
RpoE	2707457–2708032	–	16130498	9×10^{-5}	NP_951778	6	3	2
RpoH	3597560–3598414	–	16131333	2×10^{-48}	NP_951712	10	4	4
RpoN	3342358–3343791	+	16131092	6×10^{-76}	NP_952936	4	1	1

Table 1 (Continued)

Transcription factor	Location (bp) in <i>E. coli</i> genome	Strand	PID	Lowest blastp <i>E</i> -value in <i>G. sulfurreducens</i>	Accession number of the best blastp hit	No. of GI hits	No. of GI hits with $E \leq 0.1$	No. of GI hits with $E \leq 0.10^{-3}$
RpoS	2864582–2865574	–	16130648	8×10^{-67}	NP_952576	17	4	4
SoxS	4274639–4274962	–	16131888	2×10^{-5}	NP_954168	15	2	1
TorR	1056485–1057177	–	16128961	2×10^{-29}	NP_954159	108	100	84
TrpR*	4630329–4630655	+	16132210	1.3	NP_953165	10	0	0
Tus*	1682283–1683212	+	16129568	4.3	NP_954153	4	0	0
TyrR	1384744–1386285	+	16129284	3×10^{-59}	NP_951431	39	32	31

Note. Information about *E. coli* transcription factors was obtained using the *E. coli* K-12 genome search engine at the National Center for Biotechnology Information³. Abbreviations are the same as those reported by Robison et al. (1998) in the supplementary information about footprinted *E. coli* binding site matrices (Robison et al., 1998), except for the following factors that are presented by multiple subunits: FlhC and FlhD are subunits of FlhCD; HimA and HimD are subunits of IHF; and HupA and HupB are subunits of HU. All GenBank entries have the same abbreviations as those listed in the table, with the following exceptions: CarP is listed as PepA and NtrC is GlnG. Only one blastp hit with the lowest *E*-value was considered for each GI entry for *G. sulfurreducens* proteins in GenBank. An asterisk (*) marks those proteins for which the lowest *E*-value of the blastp hits exceeded 0.1. A double asterisk (**) indicates LacI, for which the blastp *E*-value was below 0.1, but which had no functional homolog identified in *G. sulfurreducens*. Accession, GenBank accession number. Shown in bold are those transcription factors for which putative orthologs were identified, determined as the best hits in the bi-directional BLAST similarity comparisons of the *E. coli* and *G. sulfurreducens* proteins. An expanded version of this table showing detailed information about the best hits in *G. sulfurreducens* and *E. coli* from bi-directional BLAST searches is provided in the online supplementary material.

2.3. Prediction of transcription factor binding sites in the genome of *G. sulfurreducens* using cross-species comparisons to established regulatory elements of *E. coli*

To compare the presence in *G. sulfurreducens* of homologs of transcription factors discussed in Section 2.1 with the presence of their predicted regulatory sites, we searched the entire genome of *G. sulfurreducens* (Methé et al., 2003) using the ScanACE software (Robison et al., 1998; Roth et al., 1998) (Table 2). This approach compared the genome sequence to 59 footprinted matrices of transcription regulatory elements from *E. coli* K-12 listed in the DPIInteract database² (Robison et al., 1998). These matrices were obtained from an online source³. In addition, if a matrix of regulatory sites reported by Robison et al. (1998) contained only four sequences or less, we verified whether a larger number of binding sites for this transcription factor was available from another database of transcription regulatory relationships of *E. coli* K-12, Regulon DB v. 4.0¹⁰ (Salgado et al., 2004). This criterion allowed us to add eight additional matrices obtained from Regulon DB with a larger representation of sequence elements. These were matrices for CysB, IlvY, MelR, ModE, NagC, OxyR, RhaS, and TorR (Table 3). In addition, we conducted separate ScanACE searches using a footprinted matrix of 6 NagC sites provided by Plumbridge (2001) and a matrix of 19 putative ArcA sites obtained from gene expression microarray analysis of the *arcA* deletion mutant (Liu and De Wulf, 2004) (Table 3).

For each possible window in the genome sequence, the ScanACE software computed the score that reflected its predicted affinity to the *E. coli* input matrix (Berg and von Hippel, 1988; Robison et al., 1998). These scores correlate with *in vitro*

binding constants (Robison et al., 1998). Similarly to previous analyses of the *E. coli* genome (Robison et al., 1998), we reported elements in *G. sulfurreducens* with scores exceeding two different cutoff values. The stringent cutoff value was the mean (μ_i) for the set of footprinted sites in *E. coli* binding site matrix *i* used in a search. The less stringent cutoff value was $\mu_i - 2\sigma_i$, where σ_i is a standard deviation (S.D.) from the mean μ_i for the set of input sites in *E. coli* matrix *i*.

Following an earlier study (Robison et al., 1998), we used multiple searches for promoter elements separated by a spacer of variable length. A separate matrix was used for each length of an internal spacer. These promoter matrices are denoted as *XL*, where *X* is the name of the transcription factor and *L* is the length of the internal spacer (Tables 2 and 3). For example, RpoD16 stands for an RpoD(σ^{70})-regulated promoter matrix with 16 bp separating the -35 and the -10 elements.

The results of the ScanACE search for putative transcription regulatory elements in *G. sulfurreducens* were compared to the results of the blastp similarity search for the presence of corresponding transcription factors. If no *G. sulfurreducens* homologs were found for a transcription factor (blastp *E*-value >0.1), any predicted binding sites for it were considered false positive hits. They were used to evaluate the limitations of the cross-species motif similarity search method.

2.4. Operon prediction

The location of putative regulatory sequence elements predicted by ScanACE was compared to the predicted boundaries of operons, ORFs, and intergenic regions in the *G. sulfurreducens* genome. Operons and ORFs were predicted using the commercial version of program FGENESB (V. Solovyev, A. Salamov, unpublished; Softberry, Inc; 2003–2004) using sequence parameters estimated from the *G. sulfurreducens*

¹⁰ <http://regulondb.ccg.unam.mx/index.html>.

Table 2
Distribution of predicted regulatory elements in the *G. sulfurreducens* genome

Transcription factor	Number of sites with scores $>\mu_i$							Number of sites with scores $>\mu_i - 2\sigma_i$							
	Total	Protein-coding			RNA	Noncoding		Total	Protein-coding			RNA	Noncoding		
		<i>N</i>	ORF	COG		GB	RNA		NC	U	<i>N</i>		ORF	COG	GB
Putative binding sites for transcription factors with homologs in <i>G. sulfurreducens</i>															
ArcA	0	0	0	0	0	0	0	139	60	32	11	1	78	67	
CpxR	3	3	0	1	0	0	0	515	344	220	70	0	171	132	
CRP	58	20	11	2	0	38	36	4014	2323	1373	406	4	1681	1405	
CspA	0	0	0	0	0	0	0	13	9	4	3	0	4	4	
CysB	0	0	0	0	0	0	0	1	1	0	0	0	0	0	
DnaA	57	29	16	4	0	28	19	117905	94207	67180	17051	301	23397	17709	
FadR	0	0	0	0	0	0	0	51	41	23	10	0	10	4	
FarR	62	24	12	2	0	38	36	1018	496	301	80	2	520	416	
Fis	123	31	10	13	0	92	79	5784	2642	1320	403	8	3134	2720	
FNR	2	1	1	0	0	1	0	100	47	18	13	1	52	40	
Fur	4	2	1	1	0	2	0	59	24	11	6	0	35	32	
GcvA	0	0	0	0	0	0	0	3	1	0	1	0	2	2	
HU	0	0	0	0	0	0	0	1	1	0	1	0	0	0	
fffF	240	36	10	8	0	204	192	75675	39429	19304	6578	150	36096	30819	
LexA	0	0	0	0	0	0	0	12	3	2	1	0	9	9	
Lrp	1117	314	102	38	0	803	684	42561	21478	11588	3702	41	21042	17513	
MalT	514	471	362	85	1	42	30	2004	1870	1469	314	1	133	92	
MetR	12	7	6	1	0	5	4	341	239	159	47	0	102	75	
NagC	0	0	0	0	0	0	0	8	1	1	0	0	7	7	
NarL	2	1	1	0	0	1	1	97	68	47	8	0	29	18	
OmpR	39	12	6	2	0	27	25	2499	1090	579	191	2	1407	1143	
PhoB	0	0	0	0	0	0	0	161	52	25	10	0	109	98	
RpoD15	457	163	77	26	0	294	255	13764	7043	3926	1242	29	6692	5523	
RpoD16	446	131	53	13	2	313	268	14782	7816	4395	1340	22	6944	5803	
RpoD17	1590	57169	237	89	0	101	871	64375	39949	24571	6926	125	24301	20059	
RpoD18	244	66	28	9	0	178	154	14756	8078	4525	1411	33	6645	5616	
RpoD19	422	157	56	23	0	265	222	28164	16777	9934	2978	88	11299	9417	
RpoH14	0	0	0	0	0	0	0	2	1	1	0	0	1	1	
RpoN	0	0	0	0	0	0	0	10	5	2	1	0	5	3	
RpoS17	229	93	45	18	0	136	117	11995	74987	4987	1472	31	4057	3308	
RpoS18	3	2	1	1	0	1	1	596	368	223	69	0	228	191	
SoxS	41	12	4	3	0	29	24	6143	3675	2194	650	15	2453	2081	
TorR	1	0	0	0	0	1	1	568	496	355	94	0	72	52	
TyrR	9	2	2	0	0	7	7	7056	4315	2664	846	11	2730	2206	
Total	5675	2156	1041	339	3	3516	3026	415172	26565	161433	45935	865	153451	126565	
False positive hits for matrices of transcription factors without a homolog in <i>G. sulfurreducens</i>															
ArgR18	2	2	0	0	0	0	0	123	56	29	9	0	67	48	
CytR	12	5	3	0	0	7	6	874	573	361	118	1	300	241	
FruR	2	0	0	0	0	2	2	41	37	26	5	0	4	4	
GalR	0	0	0	0	0	0	0	4	3	2	1	0	1	1	
GlpR	33	19	11	1	0	14	11	17089	12913	8909	2296	41	4135	3247	
H-NS	1396	809	496	162	0	587	470	22613	16684	11485	3079	74	5855	4553	
MetJ	4	1	1	0	0	3	3	635	461	313	96	0	174	135	
PurR	0	0	0	0	0	0	0	35	30	18	5	0	5	5	
Total	1449	836	511	163	0	613	492	418234	30757	21143	5609	116	10541	8234	

Note. Shown are sites predicted using 59 transcription factor binding site matrices described by Robison et al. (1998) with scores $>\mu_i - 2\sigma_i$. Transcription factor, *E. coli* transcription factor. Abbreviations for transcription factor names are described in legend to Table 1. *N*, total number of predicted sites with ScanACE scores $>\mu_i$ or $>\mu_i - 2\sigma_i$; ORF, total number of hits located in the coding regions predicted by FGENESB; COG, a subset of sites located in those ORFs that have homologs in the COG database; GB, a subset of predicted sites located in those ORFs that do not have COG database homologs but have been predicted by FGENESB to have a homologous protein in GenBank; RNA, sites within RNA genes; NC, sites predicted to be in the noncoding regions; U, a subset of predicted sites in the noncoding regions that correspond to the upstream regions of operons and singleton ORFs.

Table 3
Matrices of *E. coli* transcription regulatory sites used as input in searches of the *G. sulfurreducens* genome

Transcription factor	Number of sequences	Length of matrix (bp)	Reference	Predicted sites with scores $> \mu_i - 2\sigma_i$
Matrices for transcription factors with homologs in <i>G. sulfurreducens</i>				
Ada	3	31	Robison et al. (1998)	No
AraC	6	48	Robison et al. (1998)	No
ArcA	14	15	Robison et al. (1998)	Yes
ArcA	38	61	Regulon DB	Yes
ArcA	19	15	Liu and De Wulf (2004)	Yes
CarP	2	25	Robison et al. (1998)	No
CpxR	11	15	Robison et al. (1998)	Yes
CRP	49	22	Robison et al. (1998)	Yes
CspA	4	20	Robison et al. (1998)	Yes
CynR	2	21	Robison et al. (1998)	No
CysB	3	40	Robison et al. (1998)	Yes
CysB	7	42	RegulonDB	No
DnaA	8	15	Robison et al. (1998)	Yes
FadR	7	17	Robison et al. (1998)	Yes
FarR	4	10	Robison et al. (1998)	Yes
FhlA	3	27	Robison et al. (1998)	No
Fis	21	35	Robison et al. (1998)	Yes
FNR	14	22	Robison et al. (1998)	Yes
Fur	9	18	Robison et al. (1998)	Yes
GcvA	4	20	Robison et al. (1998)	Yes
HipB	4	30	Robison et al. (1998)	No
HU	3	16	Robison et al. (1998)	Yes
IclR	2	15	Robison et al. (1998)	No
IHF	27	48	Robison et al. (1998)	Yes
IlvY	2	27	Robison et al. (1998)	No
IlvY	4	26	Regulon DB	No
LexA	19	20	Robison et al. (1998)	Yes
Lrp	18	25	Robison et al. (1998)	Yes
MalT	10	10	Robison et al. (1998)	Yes
MarR	2	24	Robison et al. (1998)	No
MelR	2	18	Robison et al. (1998)	No
MelR	6	18	Regulon DB	Yes
MetR	8	15	Robison et al. (1998)	Yes
ModE	3	24	Robison et al. (1998)	No
ModE	6	27	Regulon DB	Yes
NagC	6	23	Robison et al. (1998)	Yes
NagC	6	23	Plumbridge (2001)	Yes
NagC	7	26	Regulon DB	Yes
NarL	11	16	Robison et al. (1998)	Yes
NarP	6	16	Robison et al. (1998)	No
NtrC	5	17	Robison et al. (1998)	No
OmpR	9	20	Robison et al. (1998)	Yes
OxyR	4	39	Robison et al. (1998)	No
OxyR	7	45	Regulon DB	Yes
PdhR	2	17	Robison et al. (1998)	No
PhoB	15	22	Robison et al. (1998)	Yes
PhoB3	4	33	Robison et al. (1998)	No
RhaS	2	50	Robison et al. (1998)	No
RhaS	3	17	Regulon DB	No
RpoD15	27	27	Robison et al. (1998)	Yes
RpoD16	48	28	Robison et al. (1998)	Yes
RpoD17	116	29	Robison et al. (1998)	Yes
RpoD18	34	30	Robison et al. (1998)	Yes
RpoD19	25	31	Robison et al. (1998)	Yes
RpoE	3	36	Robison et al. (1998)	No
RpoH13	8	49	Robison et al. (1998)	No

Table 3 (Continued)

Transcription factor	Number of sequences	Length of matrix (bp)	Reference	Predicted sites with scores $>\mu_i - 2\sigma_i$
RpoH14	7	50	Robison et al. (1998)	Yes
RpoN	6	16	Robison et al. (1998)	Yes
RpoS17	15	29	Robison et al. (1998)	Yes
RpoS18	7	30	Robison et al. (1998)	Yes
SoxS	14	35	Robison et al. (1998)	Yes
TorR	4	10	Robison et al. (1998)	Yes
TorR	6	10	Regulon DB	Yes
TyrR	17	22	Robison et al. (1998)	Yes
Matrices for transcription factors without homologs in <i>G. sulfurreducens</i>				
ArgR18 ^a	16	18	Robison et al. (1998)	Yes
ArgR3b	7	39	Robison et al. (1998)	No
CytR	5	18	Robison et al. (1998)	Yes
DeoR	3	16	Robison et al. (1998)	No
FlhCD	3	31	Robison et al. (1998)	No
FruR	12	16	Robison et al. (1998)	Yes
GalR	7	16	Robison et al. (1998)	Yes
GlpR	13	20	Robison et al. (1998)	Yes
H-NS	15	11	Robison et al. (1998)	Yes
Lac	3	21	Robison et al. (1998)	No
MetJ	15	16	Robison et al. (1998)	Yes
MetJ3	10	24	Robison et al. (1998)	No
PurR	22	26	Robison et al. (1998)	Yes
TrpR	4	24	Robison et al. (1998)	No
Tus	6	23	Robison et al. (1998)	No

Yes or No, indicates whether any *G. sulfurreducens* sites with scores $>\mu_i - 2\sigma_i$ were predicted using each matrix as ScanACE input.

^a ArgR18 represents a single binding motif, while ArgR3 corresponds to a pair of sites separated by 3 bp (according to http://arep.med.harvard.edu/ecoli_matrices/).

genome. A minimum length of 100 amino acids was used in ORF predictions. Singleton open reading frames not assigned to the operons were denoted as transcription units (Tu) by the FGENESB software. The location (predicted start and end sites) and the direction of protein-coding genes predicted by FGENESB were compared to those of the genes available from the *G. sulfurreducens* GenBank annotation¹¹ predicted by the Glimmer software (Delcher et al., 1999; Methé et al., 2003).

2.5. Distribution of predicted transcription factor binding sites among coding and noncoding regions of the *G. sulfurreducens* genome

For each predicted site with a ScanACE score above the threshold ($>\mu_i$ or $>\mu_i - 2\sigma_i$), its location relative to the predicted operons and open reading frames was identified. Any site that overlapped by at least 3 bp with an ORF predicted by FGENESB was considered to be located within a protein-coding region. Among those sites, we noted those predictions that overlapped with ORFs that had homologs in other species, as identified by the FGENESB software. These included ORFs with homologs in the clusters of orthologous groups (COG)

database (Tatusov et al., 2001), and also ORFs for which homologs in the COG database could not be found, but for which homologs in other species were found in GenBank.

Of the predicted sites that were not located within a protein-coding region, we identified those elements that overlapped by at least 3 bp with the RNA-coding regions predicted by the FGENESB software. Such sites were considered to be located in the RNA-coding regions. All remaining elements that were not classified as located in a protein-coding region or an RNA-coding region were considered to be located in the noncoding regions. We further identified those elements in the noncoding regions that were located in the upstream regions of operons or singleton open reading frames. We classified as an upstream any noncoding region located between two expressed units (operons, singleton ORFs, or RNA-coding genes), if the direction of transcription of one or both operons or singleton ORFs flanking that region was oriented away from that non-coding region. In order to evaluate the difference in nucleotide composition among different regions of the *G. sulfurreducens* genome, we computed the GC content values of the protein-coding and RNA-coding regions, and of the noncoding regions using the program geecee from the EMBOSS package (Rice et al., 2000).

We used two measures of representation of the proportion of predicted regulatory sequences in the upstream and noncoding regions described above. Each measure was computed for

¹¹ http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=genome&dopt=Protein+Table&list_uids=379.

results of searches with matrices of *E. coli* transcription factors that had homologs in *G. sulfurreducens*. For comparison purposes, these measures were also computed separately for false positive predictions found using matrices of transcription factors that had no homologs in *G. sulfurreducens*. The first measure, P , was a proportion (%) of sites predicted in the noncoding or upstream regions, e.g.:

$$P_{\mu_i, \text{upstream}} = \frac{\sum_{k=1}^{59} N_{\mu_k, \text{upstream}}}{\sum_{k=1}^{59} N_{\mu_k, \text{genome}}} \times 100\% \quad (1)$$

Here, $P_{\mu_i, \text{upstream}}$ is the proportion of sites with ScanACE scores above μ_i found in the upstream regions of operons and transcription units; $N_{\mu_k, \text{upstream}}$ is the number of predicted elements for binding site matrix k with scores above μ_i found in the upstream regions; and $N_{\mu_k, \text{genome}}$ is the number of all predicted sites for binding site matrix k with scores above μ_i found in the entire genome. Because transcription factor binding site matrices resulting in many hits affect the values of P to a greater extent than those matrices that had few hits in the genome, we also calculated the proportion of predicted elements separately for each transcription factor binding site matrix and found the average, W :

$$W_{\mu_i, \text{upstream}} = \frac{\sum_{k=1}^{59} R_{\mu_k}}{\sum_{k=1}^{59} I_{\mu_k}} \times 100\% \quad (2)$$

where $W_{\mu_i, \text{upstream}}$ is the value of W for elements with scores above μ_i found in the upstream regions of operons and transcription units; $R_{\mu_k} = N_{\mu_k, \text{upstream}}/N_{\mu_k, \text{genome}}$ if $N_{\mu_k, \text{genome}} \neq 0$, and $R_{\mu_k} = 0$ otherwise; $I_{\mu_k} = 1$ if $N_{\mu_k, \text{genome}} \neq 0$, and $I_{\mu_k} = 0$ if $N_{\mu_k, \text{genome}} = 0$. Values of P and W in all noncoding regions and/or for sites with scores $>\mu_i - 2\sigma_i$ were computed in a similar manner.

We developed Perl software, gscan.pl (by B. Yan) that lists all predicted regulatory elements in a user-specified region of the *G. sulfurreducens* genome (available from the authors upon request).

2.6. Permutation analysis

The effect of the nucleotide composition of the input *E. coli* matrices on the number of predicted regulatory sites was investigated using a permutation approach similar to that of Robison et al. (1998). Matrix sites were permuted using program SEQBOOT of the PHYLIP phylogenetic software package v. 3.64 (Felsenstein, 1989). We investigated DnaA, CRP, Fur, H-NS, and RpoD16 matrices, which had different GC content values (Robison et al., 1998) and for which different number of hits were predicted in the *G. sulfurreducens* genome (Table 2). For each matrix, the columns corresponding to sequence sites were randomly permuted using 20 replications and the “Rewrite” option of the SEQBOOT software. Each permuted replicate matrix had the overall length and nucleotide composition identical to that of the original matrix, while the sequence logo of each replicate was unique. Each replicate matrix was used as input for ScanACE to search the genome of *G. sulfurreducens*.

3. Results

3.1. *G. sulfurreducens* homologs of *E. coli* transcription factors

Table 1 provides information about the number of *G. sulfurreducens* ORFs in GenBank that show similarity to the 59 *E. coli* transcription factors from the DPInteract database. Table 1a in the online supplement¹² provides detailed information about *E. coli* and *G. sulfurreducens* ORFs identified as best BLAST hits in bi-directional protein similarity searches, their genome location, and predicted function of putative *G. sulfurreducens* transcription factors obtained from the GenBank annotation.

In searches of the *G. sulfurreducens* protein data, the use of *E. coli* regulatory proteins ArgR, CytR, DeoR, FlhC, FlhD, FruR, GalR, GlpR, H-NS, MetJ, PurR, TrpR, and Tus resulted in blastp E -values larger than 0.1, suggesting that these proteins likely do not have homologs in *G. sulfurreducens*. We verified their absence by comparing the results of the blastp search of protein data to the tblastn search results for the *G. sulfurreducens* genome. For the transcription regulators listed above, Tus had a tblastn E -value = 0.1, and all other transcription factors had $E > 0.1$ (data not shown), confirming their absence. While the remaining *E. coli* regulatory proteins used in the search had blastp hits with E -values < 0.01 (Table 1), at least one of them, LacI, likely does not have a homolog in *G. sulfurreducens*. The best blastp hit for LacI in *G. sulfurreducens* was annotated as an “ABC transporter, ATP binding protein”, with an E -value of 0.02. In addition, no transcription factor binding sites were predicted for LacI in the *G. sulfurreducens* genome (see below).

Reverse similarity searches of *E. coli* proteins using INPARANOID software identified 19 regulatory proteins as potential orthologs in *E. coli* and *G. sulfurreducens* (Table 1). These proteins were CarP, CynR, DnaA, FarR, FNR, Fur, HupB subunit of HU, both HimA and HimD subunits of IHF, LexA, ModE, NagC, NarL, PdhR, PhoB, RpoD, RpoH, RpoN, and RpoS.

Many other *E. coli* proteins listed in Table 1 likely have orthologs in *G. sulfurreducens*, even though they do not correspond to best hits in bi-directional similarity searches of *G. sulfurreducens* and *E. coli* proteins. However, due to gene duplications and deletions in both lineages, the exact identification of orthology relationships may be difficult to resolve based on reciprocal similarity searches alone. For example, the two sepa-

¹² <http://www.geobacter.org/research/genomescan/>.

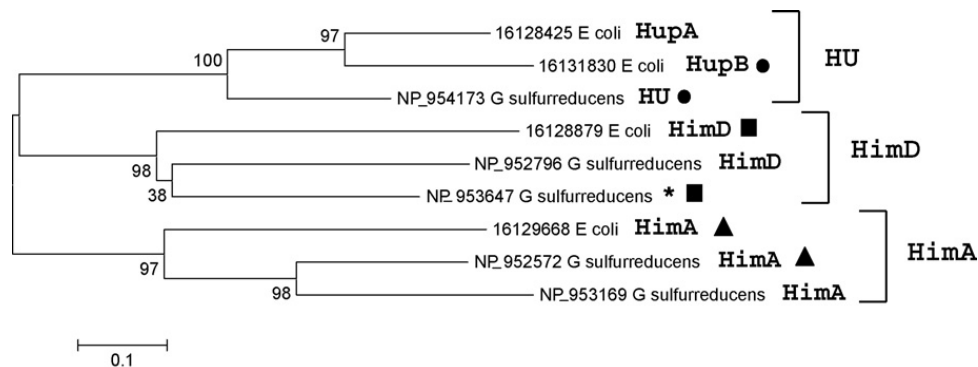


Fig. 1. Phylogenetic tree showing evolutionary relationships among protein sequences of *E. coli* and *G. sulfurreducens* subunits of HU and IHF. The tree was inferred by MEGA software v. 3.1 (Kumar et al., 2004) using the neighbor-joining method, with Poisson correction of protein distances for multiple substitutions. Numbers indicate bootstrap support (%) for each node out of 500 bootstrap replications. The scale on the bottom shows the number of amino acid substitutions per site. An asterisk (*) indicates *G. sulfurreducens* sequence (GenBank accession number NP.953647), which is likely an ortholog of *E. coli* HimD. Black circles, squares, and triangles indicate *E. coli* and *G. sulfurreducens* sequence pairs that were identified as the best blastp hits of each other in sequence similarity searches in both directions (*E. coli* proteins against all *G. sulfurreducens* proteins and vice versa; see Table 1a in the online supplement).

rate blastp searches that used both HupA and HupB subunits of *E. coli* HU as queries identified a single HU protein in *G. sulfurreducens* (GenBank accession number NP_954173) as the best blastp hit (E -values of 2×10^{-25} and 4×10^{-21} , respectively). A phylogenetic tree inferred from *E. coli* and *G. sulfurreducens* subunits of HU and their paralogs, subunits of IHF (Fig. 1), confirmed the presence of only one HU subunit (accession number NP_954173) in *G. sulfurreducens* as compared to two subunits, HupA and HupB, in *E. coli*. Each of the two *E. coli* IHF subunits, HimA and HimD, which are homologous to HU subunits, appeared to have two homologs in *G. sulfurreducens*, as a result of gene duplications. All HU and IHF protein sequences presented in Fig. 1 had a considerable degree of sequence similarity, and the E -values of the entire data set collected in blastp similarity searches using either *E. coli* or *Geobacter* proteins as queries did not exceed $E = 7 \times 10^{-10}$ (data not shown). Further studies may show how different combinations of the four IHF-like subunits in *G. sulfurreducens* participate in gene regulation.

In another example, we searched for homologs of *E. coli* RpoE, the σ^{24} subunit of RNA polymerase involved in regulation of heat shock response genes and genes with extracytoplasmic functions (Maeda et al., 2000). The best blastp hit for it in *G. sulfurreducens* was a sequence annotated in GenBank as “RNA polymerase sigma-E factor, putative” (accession number NP_951778), with $E = 9 \times 10^{-5}$. However, the best BLAST hit for NP_951778 in the reverse search of *E. coli* proteins was FecI, the σ^{19} subunit of RNA polymerase involved in the regulation of expression of genes with extracytoplasmic functions (Maeda et al., 2000),

with $E = 1 \times 10^{-10}$. The *E. coli* RpoE protein was the second best hit with $E = 1 \times 10^{-9}$ (data not shown). *G. sulfurreducens* protein NP_951778 was also the best hit when *E. coli* FecI was used as a query, with $E = 2 \times 10^{-4}$ (data not shown). A phylogenetic tree inferred from these sequences and their close homologs (Fig. 2A) suggested that *G. sulfurreducens* NP_951778 may indeed be more closely related to *E. coli* FecI than to RpoE. These close homologs included in the tree in Fig. 2A were identified from blastp searches of *E. coli* and *G. sulfurreducens*, with their E -values ≤ 0.002 . Consistent with the tree inferred, the distance between a pair of sequences corrected for multiple substitutions was the smallest between NP_951778 and *E. coli* FecI (1.48 per amino acid site) as compared to the distances between NP_951778 and *E. coli* RpoE (1.54) or between *E. coli* RpoE and FecI (1.63). Similar results were also observed when two additional divergent *G. sulfurreducens* protein sequences, with their blastp E -values > 0.1 , were added to the alignment and the tree (Fig. 2B). These sequences correspond to *G. sulfurreducens* RpoS (GenBank accession number NP_952576) (Núñez et al., 2004) and a protein annotated as “transcriptional regulator, LuxR family” (NP_953715). The similarity between NP_951778 and *E. coli* FecI may suggest that the RpoE ortholog may be absent from *G. sulfurreducens*, in agreement with the absence of predicted RpoE sites in its genome (see below). The trees presented in Fig. 2 are in full agreement with the classification of the complex evolutionary relationships within the σ^{70} family of subunits of bacterial RNA polymerase that was investigated earlier by Gruber and Gross using data from species other than *Geobacter* (Gruber and Gross, 2003).

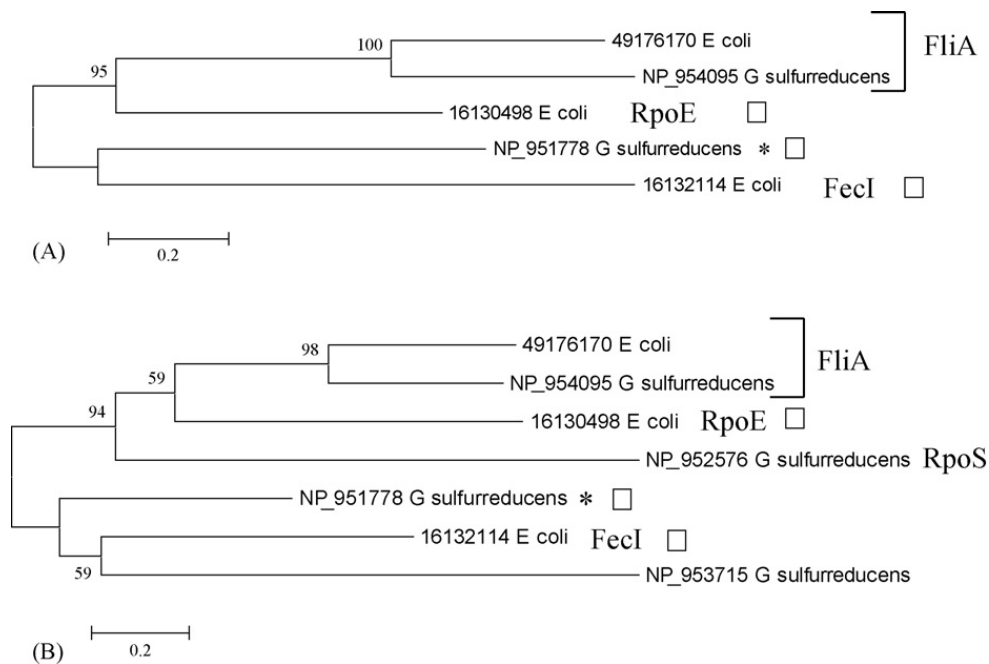


Fig. 2. Neighbor-joining trees indicating relationships of the *E. coli* RpoE protein sequence and its homologs in *E. coli* and *G. sulfurreducens*. The tree was inferred by the MEGA software with Poisson correction of protein distances for multiple substitutions. Numbers indicate bootstrap support (%) for each node. The scale on the bottom shows the number of amino acid substitutions per site. An asterisk (*) indicates *G. sulfurreducens* sequence NP_951778, which is likely more closely related to *E. coli* FecI than to RpoE. White squares indicate the *E. coli* RpoE protein (GenBank accession number 16030498), its best blastp hit in *G. sulfurreducens* (NP_981778), and the *E. coli* FecI protein (16132114). (A) Phylogenetic tree inferred from close homologs with blastp *E*-values not exceeding 0.002. (B) Phylogenetic tree inferred after adding more distant homologs with blastp *E*-values exceeding 0.1.

3.2. Operon and gene predictions

The predicted operon locations and gene predictions in the *G. sulfurreducens* genome by the FGENESB program are provided in the online supplement for this report. All analyses presented here are based on the reference *G. sulfurreducens* operon annotation of February 25, 2004. Later annotation updates confirmed the predicted operon organization of the *G. sulfurreducens* genome as compared to the reference annotation of February 25, 2004 discussed in this report (unpublished data). This version of annotation was recently employed in the microarray analysis of the *G. sulfurreducens* RpoS deletion mutant (Yan et al., 2006). The structure of a number of the operons presented in this annotation has been confirmed experimentally (Núñez et al., 2004; Leang and Lovley, 2005).

The FGENESB program predicted 766 operons and 731 singleton ORFs not assigned to any operon (the latter referred to by the software as transcription units). The total number of protein-coding genes predicted by the program was 3501. The online supplementary material provides the detailed information about 2587 of these genes that were completely identical (the same start, end, and the genome strand orientation) to the 3466 genes

available from the *G. sulfurreducens* GenBank annotation as predicted by the Glimmer software (Delcher et al., 1999) as part of the sequencing and annotation the *G. sulfurreducens* genome (Méthé et al., 2003). In addition, for 348 more genes the FGENESB software predicted their start and direction to be identical to those provided by the GenBank annotation, but the end positions of these genes were predicted differently by the two annotations. The total number of genes for which their start position and strand direction were predicted to be identical to that available from GenBank was 2935, which represents 83.8% of the 3501 protein-coding genes predicted by FGENESB and 84.7% of the 3466 genes predicted by Glimmer (Méthé et al., 2003). In the present study, we consider the classification of genome regions as protein-coding, RNA-coding, and noncoding according to the predictions made by the FGENESB software.

The total combined length of the protein-coding regions in *G. sulfurreducens* predicted by the FGENESB software was 3430729 bp, or 89.9% of the entire genome. This is comparable to 87.8% of the *E. coli* genome that corresponds to the protein-coding genes (Blattner et al., 1997). The predicted ORFs for which homologs were found in the COG database had a combined length of

2564787 bp. The length of ORFs for which no homologs were found in the COG database, but for which homologs were found in GenBank, was 631074 bp. Additional ORFs were predicted by the FGENESB software that had no known homologs. The combined length of predicted RNA-coding genes was 8078 bp, or 0.2% of the genome. The combined length of the regions outside predicted protein-coding or RNA-coding genes was 375332 bp, or 9.8% of the entire genome. Of these regions, 263149 bp, or 6.9% of the entire genome, were located upstream of predicted operons and transcription units. The total of 1095 upstream regions in the genome included 108 regions with length <100 bp; 500 regions of length ≥ 100 bp, but <200 bp; 242 regions ≥ 200 bp in length, but <300 bp; and 245 additional upstream regions with length ≥ 300 bp.

3.3. Putative binding sites predicted for transcription factors with homologs in *G. sulfurreducens*

All elements predicted in the genome of *G. sulfurreducens* with scores above mean (μ_i) and $\mu_i - 2\sigma_i$ are provided in the online supplement. Table 2 shows the results of the search of the *G. sulfurreducens* genome using matrices of aligned footprinted *E. coli* binding sequences of Robison et al. (1998). In the present Section, we discuss putative binding sites for *E. coli* transcription factors with predicted orthologs or homologs in *G. sulfurreducens*. Some of these hits may be true transcription regulatory sites, while others may include false positive predictions.

For 19 transcription factors that likely had orthologs or homologs in *G. sulfurreducens*, no respective sites were found for searches with matrices of Robison et al. (1998) with scores exceeding the lowest cutoff level, $\mu_i - 2\sigma_i$. These regulators were Ada, AraC, CarP, CynR, CysB, FhlA, HipB, IclR, IlvY, MarR, MelR, ModE, NarP, NtrC, OxyR, PdhR, PhoB3, RhaS, RpoE, and RpoH13. When additional matrices from Regulon DB with a larger number of footprinted sites were used (Table 3), no elements were found with scores $>\mu_i - 2\sigma_i$ for IlvY and RhaS, consistent with the search using matrices of Robison et al. (1998). The use of Regulon DB ModE and OxyR matrices did not identify any sites with scores $>\mu_i$, but it predicted 80 sites for ModE and 8 sites for OxyR with scores $>\mu_i - 2\sigma_i$ (see the online supplement and the description of ModE sites in Section 4, below). The use of Regulon DB MelR matrix provided 1 site with a score $>\mu_i$ and 32 sites with scores $>\mu_i - 2\sigma_i$.

Results of searches with matrices of binding site sequences reported by Robison et al. (1998) with scores

$>\mu_i - 2\sigma_i$ are listed in Table 2. Eleven matrices resulted in hits with scores above $\mu_i - 2\sigma_i$, but below μ_i . These were predicted binding sites for ArcA, CspA, CysB, FadR, GcvA, HU, LexA, NagC, PhoB, RpoH14, and RpoN (Table 2). For CysB, however, only one putative site in a coding region was found with a score between $\mu_i - 2\sigma_i$ and μ_i (Table 2). No elements with scores $>\mu_i - 2\sigma_i$ were predicted when the CysB matrix from Regulon DB was used.

For NagC, the number (34) of elements with scores $>\mu_i - 2\sigma_i$ predicted using the Regulon DB matrix or that (32) for the matrix of Plumbridge (2001) was comparable to the number (8) of sites predicted using the matrix of Robison et al. (1998). Neither search identified any elements with a score $>\mu_i$.

For ArcA, the search using an *E. coli* matrix compiled from microarray data (Liu and De Wulf, 2004) (Table 3) returned 4 putative sites with scores $>\mu_i$ and 196 hits with scores $>\mu_i - 2\sigma_i$, compared to 139 sites with scores $>\mu_i - 2\sigma_i$ identified when using the matrix of Robison et al. (1998). When the predicted ArcA sites with scores $>\mu_i - 2\sigma_i$ were compared, 131 nonoverlapping sites predicted using the microarray-based matrix overlapped by at least 6 bp with elements predicted using matrices of Robison et al. (1998). This suggests that both searches identified mostly the same ArcA sites.

Twenty-three binding site matrices resulted in hits with scores $>\mu_i$, the most restrictive cutoff. These were matrices for CpxR, CRP, DnaA, FarR, Fis, FNR, Fur, IHF, Lrp, MalT, MetR, NarL, OmpR, RpoD15, RpoD16, RpoD17, RpoD18, RpoD19, RpoS17, RpoS18, SoxS, TorR, and TyrR (Table 2). When the Regulon DB TorR matrix with a larger number of input sites was used, a larger number of hits were predicted (51 sites with scores $>\mu_i$ and 4736 with scores $>\mu_i - 2\sigma_i$ when using the Regulon DB matrix, compared to 1 and 568 sites, respectively, for matrices of Robison et al., 1998).

As discussed in Section 3.1, four subunits of IHF and one subunit of HU (Laine et al., 1980; Weglenska et al., 1996) are present in *G. sulfurreducens* (Table 1; Fig. 1). While 240 elements with scores $>\mu_i$ and 75675 elements with scores $>\mu_i - 2\sigma_i$ were found for IHF, no sites with scores $>\mu_i$ and only one hit with a score $>\mu_i - 2\sigma_i$, located in a protein-coding region, was found for HU (Table 2). The absence of predicted HU binding sites is consistent with the observation by other authors that HU binds to DNA in a sequence nonspecific manner, recognizing curved DNA (Azam and Ishihama, 1999). Such HU binding sites would not be detected by sequence similarity searches.

3.4. Distribution among the noncoding and the coding regions of putative binding sites for transcription factors with homologs present in *G. sulfurreducens*

When searching with binding site matrices that correspond to transcription factors that had putative homologs in *G. sulfurreducens*, we identified the total of 5675 putative elements in the *G. sulfurreducens* genome with scores $>\mu_i$ (Table 2). They included 2156 sites, or 38.0% of the total number of hits with scores $>\mu_i$ in protein-coding regions and three hits in RNA-coding regions. Even though noncoding regions represented 9.8% of the genome length, 3516 sites with scores $>\mu_i$ were found in noncoding regions. They correspond to the proportion $P_{\mu_i, \text{noncoding}} = 62.0\%$ of the total number of predicted elements with scores $>\mu_i$. Of the sites in the noncoding regions, 3026 were located in the upstream regions of operons and transcription units. This corresponds to a proportion of $P_{\mu_i, \text{upstream}} = 53.3\%$, while the length of the upstream regions of the operons and singleton ORFs is only 6.9% of the total length of the genome. We also computed the weighted average proportion, W , of the number of predicted sites with scores $>\mu_i$ that involved separate comparisons for each of the binding site matrices for different transcription factors. This measure was also very high in the noncoding regions: $W_{\mu_i, \text{noncoding}} = 58.8\%$ and $W_{\mu_i, \text{upstream}} = 48.7\%$.

The high prevalence of predicted hits in the upstream regions of operons and transcription units suggests that a number of these sites may be true regulatory elements, because transcription regulatory elements are more likely to occur in the upstream regions than within ORFs. However, some functional transcriptional regulatory elements in *G. sulfurreducens* occur in the coding and intergenic regions within operons. For example, elements within protein-coding regions regulate the expression of the *omcB* and the *omcC* operons (Leang and Lovley, 2005). Therefore, the location of a predicted element in an upstream region cannot serve as the sole criterion of its possible functional role.

Similar to results obtained by other authors for *E. coli* (Robison et al., 1998), the predicted sites satisfying less stringent criteria of scores $>\mu_i - 2\sigma_i$ were much more abundant than those satisfying the scores $>\mu_i$. The total number of elements with scores $>\mu_i - 2\sigma_i$ that corresponded to search matrices of transcription factors with likely homologs in *G. sulfurreducens* was 415172 (Table 2). The use of the less stringent cutoff value identified many hits that could not be detected with the stringent cutoff value, μ_i , in *G. sulfurreducens*. However, many of these sites may be false positive predictions.

Only 153451 of these sites were located in the noncoding regions ($P_{\mu_i - 2\sigma_i, \text{noncoding}} = 37.0\%$; $W_{\mu_i - 2\sigma_i, \text{noncoding}} = 41.7\%$), and 126565 motifs ($P_{\mu_i - 2\sigma_i, \text{upstream}} = 30.5\%$; $W_{\mu_i - 2\sigma_i, \text{upstream}} = 35.4\%$) were found in the upstream regions, similar to proportions obtained by other authors for *E. coli* data (Robison et al., 1998).

The distributions of the binding sites predicted using additional matrices from Regulon DB and from the ArcA microarray data were similar to the predictions obtained using matrices of Robison et al. (1998) (data not shown).

The online supplementary material provides the list and location of the predicted elements within the upstream regions of each operon and transcription unit in the *G. sulfurreducens* genome.

Some elements in the upstream regions may have been predicted due to a difference in nucleotide composition between the noncoding and coding regions. The GC content is unevenly distributed in the *E. coli* genome and depends on the biological properties of the region (Blattner et al., 1997). Similar to *E. coli*, the GC content is lower in the noncoding regions of the *G. sulfurreducens* genome: its average value is 62% for the combined protein-coding regions, 58% for the combined RNA-coding regions, and 53% for the combined noncoding regions. The latter includes the upstream regions of operons and transcription units, for which the average value is 52%. However, an earlier study (Robison et al., 1998) that used permutation of binding site matrix columns to search the *E. coli* genome suggested that sites identified by ScanACE in the upstream regions are predicted based on true specificity of search matrices rather than due to the difference in the nucleotide composition between the upstream and the noncoding regions. Below we discuss the results of permutation analyses that investigated the effect of the nucleotide composition bias on the genome location of predicted binding sites in *G. sulfurreducens*.

3.5. False positive predictions of binding sites for transcription factors with no homologs in *G. sulfurreducens*

Any binding sites predicted for transcription factors missing from *G. sulfurreducens* likely represent false positive predictions, as no DNA binding by their transcription factors can occur, and therefore there is no selection pressure to preserve these sites in the genome. For several missing transcriptional regulators, the results of the ScanACE motif search were consistent with their absence. No hits with scores $>\mu_i - 2\sigma_i$ were predicted for DeoR, FlhCD, TrpR, and Tus, confirming the absence of these regulators (Table 2). Notably, both subunits,

FlhC and FlhD, of FlhCD (Prüß et al., 2003), were absent from *G. sulfurreducens*, consistent with the absence of FlhCD binding sites. Additionally, no hits were found for LacI, confirming its absence from *G. sulfurreducens*, as described above.

For a number of transcription factors missing from *G. sulfurreducens*, multiple false positive elements were predicted with scores $>\mu_i$. These included ArgR (predicted sites for the ArgR18 matrix had scores $>\mu_i$, whereas for the ArgR3 matrix no hits with scores $>\mu_i - 2\sigma_i$ were found), CytR, FruR, GlpR, H-NS, and MetJ (predicted sites for the MetJ matrix had scores $>\mu_i$, while no sites with scores $>\mu_i - 2\sigma_i$ were found for the MetJ3 matrix). Interestingly, while the best blastp hit for *E. coli* MetJ in protein sequence similarity searches had an *E*-value of 0.41 in *G. sulfurreducens* (GenBank accession NP_953038), this protein was annotated in GenBank as a “sigma-54 dependent DNA-binding response regulator”. The high *E*-value indicates a significant sequence divergence between *E. coli* MetJ and *G. sulfurreducens* NP_953038, suggesting that a true ortholog of MetJ may be absent from *G. sulfurreducens*, and therefore elements identified by ScanACE for the MetJ matrix are likely false positive hits.

Of particular interest is H-NS, for which 1396 putative recognition sites were predicted with scores $>\mu_i$, and 22613 elements had scores $>\mu_i - 2\sigma_i$. Of the elements with scores $>\mu_i$, 587 were in the noncoding regions, including 470 elements in the upstream regions. BLAST similarity searches clearly show the absence of the H-NS homolog from *G. sulfurreducens* (Table 1). An earlier study predicted over 5700 H-NS elements with scores $>\mu_i$ and over 63,500 elements with scores $>\mu_i - 2\sigma_i$ in *E. coli*, noting the loose sequence specificity of H-NS for its binding sites (Robison et al., 1998). H-NS recognizes curved DNA rather than specific sequence sites (Azam and Ishihama, 1999). Our results suggest that predicted H-NS elements in the genome of *G. sulfurreducens* may not be true H-NS sites. While it is possible that they are recognized by a transcription factor other than H-NS, it is most likely that the footprinted *E. coli* H-NS binding site matrix contains very divergent sequences, resulting in a large number of false positive hits.

The total number of false positive hits with scores $>\mu_i$ found for transcription factors absent from *G. sulfurreducens* was 1449 (Table 2). This included 836 sites in the protein-coding regions and 613 sites in the noncoding regions ($P_{\mu_i, \text{noncoding}} = 42.3\%$; $W_{\mu_i, \text{noncoding}} = 53.0\%$), including 492 predicted sites in the upstream regions ($P_{\mu_i, \text{upstream}} = 34.0\%$; $W_{\mu_i, \text{upstream}} = 48.7\%$). No elements were predicted in the RNA-coding genes. As expected, the proportion of hits in the noncoding

and the upstream regions predicted for transcription factors absent from *G. sulfurreducens* was lower than the corresponding proportion of predicted sites for transcription factors with homologs in *G. sulfurreducens* ($P_{\mu_i, \text{noncoding}} = 62.0\%$; $W_{\mu_i, \text{noncoding}} = 58.8\%$; $P_{\mu_i, \text{upstream}} = 53.3\%$; $W_{\mu_i, \text{upstream}} = 48.7\%$). The similarity between the values of *W* for the false positive hits and the likely true binding sites was influenced by a small number of false positive hits for FruR and MetJ, which affected the values of *W* (two and four sites, respectively, with nearly all sites predicted in the upstream regions, as presented in Table 2).

The total number of predicted false positive hits satisfying less stringent criteria of ScanACE scores being $>\mu_i - 2\sigma_i$ was 41414 (Table 2). In addition to ArgR18, CytR, FruR, GlpR, H-NS, and MetJ, which also had hits with scores $>\mu_i$, described above, two other proteins absent from *G. sulfurreducens*, GalR and PurR, had predicted sites with scores $>\mu_i - 2\sigma_i$, but not exceeding μ_i . The 41414 false positive hits with scores $>\mu_i - 2\sigma_i$ included 30757 sites in protein-coding regions, 116 in RNA-coding regions, and 10541 in the noncoding regions, including 8234 sites in the upstream regions ($P_{\mu_i - 2\sigma_i, \text{noncoding}} = 25.5\%$; $W_{\mu_i - 2\sigma_i, \text{noncoding}} = 26.9\%$; $P_{\mu_i - 2\sigma_i, \text{upstream}} = 19.9\%$; $W_{\mu_i - 2\sigma_i, \text{upstream}} = 22.0\%$).

It is apparent that a larger number of false positive hits are predicted in the upstream regions than expected by chance. The difference in nucleotide composition of the upstream regions and the coding regions of the genome likely has some effect on the distribution of predicted sites. However, the values of *P* and *W* for false positive hits are considerably lower than the corresponding values for the putative binding sites of transcription factors with homologs in *G. sulfurreducens*, which were $P_{\mu_i - 2\sigma_i, \text{noncoding}} = 37.0\%$; $W_{\mu_i - 2\sigma_i, \text{noncoding}} = 41.7\%$; $P_{\mu_i - 2\sigma_i, \text{upstream}} = 30.5\%$; $W_{\mu_i - 2\sigma_i, \text{upstream}} = 35.4\%$.

The GC content is within the same range for the input *E. coli* binding site matrices of transcription factors with homologs in *G. sulfurreducens* and for matrices corresponding to transcription regulators without such homologs (Robison et al., 1998). Therefore, the differences in the distribution of the predicted hits cannot be explained by nucleotide composition alone, suggesting that some of the predicted hits for transcription factors with likely homologs in *Geobacter* may be functional regulatory sites.

3.6. Permutation analysis

The role of the difference in nucleotide composition between the coding and the noncoding regions was investigated using ScanACE searches of the *G. sulfurre-*

Table 4

Comparison of predicted regulatory elements using original *E. coli* binding site matrices and their permuted replicates

Matrix	Predicted sites with scores $>\mu_i$			Predicted sites with scores $>\mu_i - 2\sigma_i$		
	Total	Proportion in the noncoding regions		Total	Proportion in the noncoding regions	
	<i>N</i>	PNC (%)	PU (%)	<i>N</i>	PNC (%)	PU (%)
Transcription factors with homologs in <i>G. sulfurreducens</i>						
CRP	56.9 ± 10.5 (58)	56.7 ± 9.5 (65.5)	48.0 ± 9.4 (62.1)	4078.2 ± 441.4 (4014)	40.7 ± 3.9(41.9)	33.4 ± 3.3(35.0)
DnaA	32.9 ± 9.9 (57)	34.8 ± 10.8 (49.1)	28.6 ± 9.9 (33.3)	115683.8 ± 7216.9 (117905)	17.8 ± 1.1(19.8)	13.6 ± 0.8(15.0)
Fur	2.0 ± 1.2 (4)	58.8 ± 41.2 (50.0)	46.1 ± 43.9 (0.0)	48.3 ± 10.0 (59)	65.1 ± 6.8(59.3)	55.1 ± 7.4 (54.2)
RpoD16	444.3 ± 38.5 (446)	66.4 ± 2.5 (70.2)	58.3 ± 2.7 (60.1)	14411.8 ± 922.8 (14782)	48.4 ± 1.9(47.0)	40.7 ± 1.7(39.3)
Transcription factor without a homolog in <i>G. sulfurreducens</i>						
H-NS	1426.4 ± 348.5 (1396)	36.5 ± 6.4 (42.0)	29.3 ± 5.3 (33.7)	21787.8 ± 3933.4 (22613)	24.8 ± 3.6(25.9)	19.3 ± 2.9 (20.0)

Note. Values are presented as mean ± standard deviation, computed using predictions from ScanACE searches with 20 permuted replicates of each matrix. Numbers in parenthesis are presented for comparison purposes, and they show the values computed using the original non-permuted *E. coli* matrices (Table 2). *N*, the total number of predicted sites with scores $>\mu_i$ or $>\mu_i - 2\sigma_i$; PNC, proportion (%) of sites (out of the total number *N*) that were predicted to be in the noncoding regions; PU, proportion (%) of the sites (out of the total number *N*) predicted to be in the upstream regions.

ducens genome with 20 permuted replicates of *E. coli* binding site matrices for CRP, DnaA, Fur, RpoD16. All these transcription factors had homologs in *G. sulfurreducens* (Table 1). In each replicate, columns corresponding to sequence sites were randomly permuted, while the overall nucleotide composition and the length of the permuted matrix did not change. As shown in Table 4, when the permutation procedure was applied, the average number of predicted elements with scores $>\mu_i$ slightly decreased for the permuted CRP, DnaA, Fur, and RpoD16 matrices compared to the original matrices. The average number of sites with scores $>\mu_i - 2\sigma_i$ also decreased for the permuted DnaA, Fur, and RpoD16 matrices compared to the original matrices, but this number increased for the permuted CRP matrix. The changes in the number of predicted *G. sulfurreducens* elements for the permuted CRP matrix were in agreement with the trends observed in *E. coli* (Robison et al., 1998). Similar to the results for *E. coli*, we found that the proportion of the sites with scores $>\mu_i$ in the upstream and noncoding regions sharply decreased when the columns of the CRP matrices were permuted (56.7% versus 65.5%, respectively, on average, for the sites predicted in the noncoding regions using the permuted CRP matrix as compared to the original matrix; and 48.0% versus 62.1% for the hits in the upstream regions). Such proportions were also decreased, although to a lesser extent, for the elements scoring $>\mu_i - 2\sigma_i$: 40.7% versus 41.9% for the hits in the noncoding regions, and 33.4% versus 35.0% for the upstream regions. Similarly, the proportions of hits with scores $>\mu_i$ in the noncoding and the upstream regions were sharply decreased for the permuted DnaA and RpoD16 matrices. For example, the proportion of such

sites in the noncoding regions was on average 34.8% for the permuted replicates of the DnaA matrix compared to 49.1% for the original matrix. These proportions were also lower for the permuted replicates of the DnaA matrix when considering scores $>\mu_i - 2\sigma_i$ (Table 4). In contrast, the proportion of elements with scores $>\mu_i - 2\sigma_i$ in the upstream and the noncoding regions slightly increased for the permuted replicates of the RpoD16 matrix, suggesting that the use of such a low threshold for this matrix may yield many nonspecific hits.

The GC content of CRP, DnaA, and RpoD16 matrices is in a similar range of 37.0%, 36.7%, and 34.9%, respectively (Robison et al., 1998). These examples suggest that shuffling of the columns of these matrices decreased the specificity of binding site predictions, in particular when selecting sites with scores $>\mu_i$. However, two additional examples, Fur and H-NS, suggested that the permutation procedure may have some limitations. Fur, ferric uptake repressor, has an ortholog in *G. sulfurreducens* (Rodionov et al., 2004). In *E. coli*, the consensus of its binding sites is a palindromic sequence, and the 18 bp long Fur matrix used in the searches has a very low GC content of 23.4% (Robison et al., 1998; Panina et al., 2001). As a result, permutations of the columns of the Fur matrix frequently lead to a replacement of A or T with the same nucleotide, resulting in many permuted replicates that contain poly-A and poly-T stretches and are very similar in sequence to the original matrix (data not shown). After the permutation of the Fur matrix, the proportion of sites predicted in the noncoding and the upstream regions increased both for sites with scores $>\mu_i$, and for sites with scores $>\mu_i - 2\sigma_i$. As discussed below, a number of elements predicted using the origi-

nal Fur matrix may be functional Fur sites. The increase in the proportion of the predicted hits in the noncoding and the upstream regions after permuting the columns of the Fur matrix is likely due to the limitations of the permutation procedure, which replaces nucleotides in a short matrix in a nonrandom manner.

To further investigate the effects of the permutation procedure, it was applied to the H-NS matrix, which corresponds to a transcription factor without a homolog in *G. sulfurreducens*, and therefore predicted H-NS sites are likely false positive hits. Therefore, permutation of the columns of that matrix should not influence the overall number of predicted sites or their distribution between the coding and the noncoding regions. Indeed, the use of permuted H-NS replicates resulted in a number of sites similar to those predicted for the original H-NS matrix (Table 4). The mean number of predicted elements with scores $>\mu_i$ among the 20 replicates was 1426.4, which is slightly higher than 1396 predicted sites for the original H-NS matrix. The mean number of sites with scores $>\mu_i - 2\sigma_i$ was 21787.8, compared to 22613 sites for the original matrix. Interestingly, the average values (computed from 20 permuted replicates of the H-NS matrix) of proportions of predicted hits in the noncoding and the upstream regions were lower than similar proportions for the original H-NS matrix (Table 4). For example, the average of 36.5% of hits with scores $>\mu_i$ were in the noncoding regions when 20 permuted replicates of the H-NS matrix were used, compared to 42.0% for the original matrix. For the sites in the upstream regions with scores $>\mu_i$, these values were 29.3% and 33.7%, respectively. On average 24.8% of sites with scores $>\mu_i - 2\sigma_i$ were in the noncoding regions when using 20 permuted replicates, as compared to 25.9% for the original H-NS matrix. In the upstream regions, these proportions were 19.3% and 20.0%, respectively. Similar to the Fur matrix discussed above, the H-NS matrix is short (11 bp) and has low GC content of 31.5% (Robison et al., 1998). Therefore, the shuffling of its columns may have been nonrandom and resulted in a frequent replacement of a nucleotide from the original matrix with the same nucleotide in the permuted matrix. The Fur and H-NS examples show the limitations of the permutation procedure when applied to short sequences with a bias in their nucleotide composition. Evidence for functional significance of predicted sites for transcription factors present in *G. sulfurreducens* may come from additional considerations, e.g., their cross-species conservation in the upstream regions of functionally important genes, or upstream of co-regulated operons under the control of a transcription factor of interest.

4. Discussion

Regulatory sites predicted in the upstream regions provide an insight into the pathways of operon regulation in *G. sulfurreducens*. Our results indicate that the use of known footprinted binding site matrices from a divergent bacterial species is a useful approach to identify candidate genome regions affecting gene expression in an organism of interest. The knowledge obtained from understanding the individual elements involved in gene regulation can be used for future analyses that would reconstruct gene regulatory networks, a key for understanding cellular behavior, and for developing cellular models. However, it is necessary to verify that a transcription factor of interest is not absent from the organism under study. Additionally, the predictions of regulatory elements may be affected by their cross-species divergence and by the genetic polymorphisms of their sequences within each species.

Similarity of sequence matrices from one species to sequence elements in a genome of a divergent species may be influenced by the phylogenetic divergence of both transcription factors and their binding sites, especially if multiple homologs of a particular transcription factor exist. For example, the results of our search did not discriminate between the predicted $-35/-10$ promoter elements that correspond to different sigma factors, e.g., RpoS and RpoD promoter elements recognize similar sequences (Lee and Gralla, 2001; Hengge-Aronis, 2002; Lacour et al., 2003; Lacour and Landini, 2004; Weber et al., 2005; Eggers et al., 2006; Yan et al., 2006). In a number of cases, ScanACE searches using *E. coli* RpoS and RpoD matrices recognized essentially the same elements in the *G. sulfurreducens* genome. Housekeeping RpoD promoters may determine expression of many operons, while other sigma factors from the σ^{70} family are competitively involved in binding to the core RNA polymerase, initiating transcription of specialized sets of genes under specific conditions (Maeda et al., 2000). Additional transcription modulators and repressors other than sigma factors may affect expression of certain genes and operons under specific conditions, and therefore additional regulatory sites may be differentially present in certain operons (Hengge-Aronis, 2000; Maeda et al., 2000). The challenges in distinguishing among promoters regulated by different sigma factors using computational methods have been documented previously (Huerta and Collado-Vides, 2003). Our recent use of microarray data and primer extension analyses in *G. sulfurreducens* suggested initial confirmation of a number of RpoS-regulated promoters as compared to

promoters regulated by other sigma factors (Yan et al., 2006).

For many binding site matrices that returned positive hits, the number of predicted elements was in a similar range (<10, <100, <1000, or ≥ 1000) for the *G. sulfurreducens* genome search presented here and for a similar search in *E. coli* (Robison et al., 1998) at comparable cutoff level for ScanACE scores. However, in general fewer elements were predicted for *G. sulfurreducens* than for *E. coli*. This may be related to a smaller size of the *G. sulfurreducens* genome (3814139 bp) (Méthé et al., 2003) as compared to that of *E. coli* (4639221 bp) (Blattner et al., 1997), and to the differences in regulatory circuitry and in the number of protein copies in the cells between these species. Another very plausible explanation is the sequence divergence of *G. sulfurreducens* regulatory elements from those in *E. coli*. For example, 220 hits with scores $>\mu_i$ were found in *E. coli* for the CRP matrix (Robison et al., 1998), whereas our search using the same matrix found 58 such hits in *G. sulfurreducens* (Table 2). Previous studies have shown that many *E. coli* CRP sites may not be detected by similarity searches (Djordjevic et al., 2003; Brown and Callan, 2004). Two CRP-FNR family members, referred to as Flp, or FNR-like proteins, have been found in *G. sulfurreducens* (Esteve-Nunez et al., 2004) (see also Table 1). Their GenBank accession numbers are NP_954461 and NP_953041. NP_954461, also denoted HcpR, or GSU3421, has been suggested to recognize a *G. sulfurreducens* motif with a third position of its consensus sequence different from that for CRP (Rodionov et al., 2004; Rodionov et al., 2005). Future investigation will determine which of the sequence elements in *G. sulfurreducens* predicted using CRP and FNR matrices (Table 2 and the online supplement) may be recognized by regulators NP_954461 and NP_953041.

Another transcription factor with a smaller number of predicted binding sites in *G. sulfurreducens* than in *E. coli* is Fur, for which 36 elements with ScanACE scores $>\mu_i$ were found in *E. coli* (Robison et al., 1998), while 4 such elements were found in *G. sulfurreducens* (Table 2). Fur is a functional transcriptional regulator in *G. sulfurreducens* (O'Neil et al., 2004). When using the cutoff value of $\mu_i - 2\sigma_i$, we identified 59 putative Fur sites in *G. sulfurreducens* (see the online supplement for their genome locations). They included three of the four Fur boxes identified in an earlier study of sequence conservation among δ -*Proteobacteria* in the upstream regions of the *fur*, *X-feoA-feoA-feoB2*, and *genX-genY* operons (Rodionov et al., 2004), with genome positions starting at 1507886, 3589565, and 3591220, respectively. These results suggest that the score cutoff needs to be low-

ered below μ_i in a cross-species comparison in order to identify a broad range of potential Fur binding sites.

The ability of a binding site matrix to recognize similar sites in a genome is affected not only by its sequence divergence from the genome under study, but also by its length, the number of sequences in the matrix, and the degree of their conservation (Robison et al., 1998). In the searches reported here, no hits were frequently found when using transcription factor binding site matrices that had a small number of sequences (six or less) (Table 3). The effect of sequence length was less obvious, as search results either returned no hits or returned positive results both for input matrices as short as 10 bp and as long as 50 bp. A recent study (Fogel et al., 2005) of eukaryotic transcription factor binding site matrices from the TRANSFAC database suggested the important role of the sequence length in matrix specificity, suggesting that 21–43% of the entire matrix length represented a matrix core affecting its specificity. It is very likely that, similar to eukaryotic TFBS matrices, the sequence length plays an important role in finding transcription factor-specific hits rather than false positive results.

The role of input sequences is illustrated by the molybdate-responsive transcription factor, ModE, which is present in *G. sulfurreducens* (GenBank accession number NP_954006) (Rodionov et al., 2004) (Table 1). A site similar to the *E. coli* ModE binding site consensus has been found earlier upstream of the *G. sulfurreducens* *modABC* operon (Rodionov et al., 2004), and it is also present in other species of *Proteobacteria*, green sulfur bacteria, and *Archaea* (Studholme and Pau, 2003). When we searched the *G. sulfurreducens* genome with the *E. coli* ModE matrix (three input sequences; 24 bp long; $\mu_i = 33.72$, $\sigma_i = 1.18$) of Robison et al. (1998), without using a score cutoff, this site had the highest score of 18.4939 of all putative ModE hits, but it did not exceed the $\mu_i - 2\sigma_i$ threshold (data not shown), suggesting that the $\mu_i - 2\sigma_i$ cutoff was too stringent for this search. When the ModE matrix from Regulon DB was used as ScanACE input (six sequences; 27 bp long; $\mu_i = 26.95$, $\sigma_i = 7.03$), 80 hits were found with scores $>\mu_i - 2\sigma_i$. Among them, the site upstream of the *modABC* operon (operon 1300) (Studholme and Pau, 2003; Rodionov et al., 2004) had the fifth rank among scores, with a score of 18.2764, as presented in the online supplement. That site (with a 3 bp window shift) was also predicted with the rank of 47 and the score of 13.8505. Other elements with scores $>\mu_i - 2\sigma_i$ found using the ModE matrix from Regulon DB were, for example, a site with a score of 18.0738 (genome positions 3757162–3757188), upstream of operon 1481 that contains a gene encod-

ing a putative Fe-S oxidoreductase, family 2 (COG1032) (accession number NP_954455). Its product, annotated as “radical SAM domain protein/B12 binding domain protein”, was determined by the blastp searches to share a distant homology with the products of the *moaA* gene in *Archaeoglobus fulgidus* and *Sulfolobus tokodaii*, which encodes the molybdenum cofactor biosynthesis protein A (data not shown). The operon containing the *moaA* gene has a ModE-regulated promoter in *E. coli* (McNicholas et al., 1997). While NP_954455 appears to be divergent from the *moaA* gene product, several ModE sites have been found upstream of *E. coli* Fe-S oxidoreductases genes (Studholme and Pau, 2003), and therefore the predicted ModE element in *G. sulfurreducens* is likely a functional site. An element with a score of 16.9099 (positions 3748873–3748899) was identified upstream of operon 1476 that encodes a putative membrane protein (NP_954447) with homology to permeases of the drug/metabolite transporter (DMT) family (COG0697). A ModE site was previously detected upstream of a homolog of this gene in *Yersinia pestis* (Studholme and Pau, 2003). A likely ModE element with a score of 18.811 in positions 2393719–2393745 is located upstream of glycine (CCC) tRNA, while an earlier study (Studholme and Pau, 2003) predicted a ModE site upstream of the *E. coli glyQ* gene that encodes the alpha subunit of glycine tRNA synthetase. These examples suggest likely functional roles of many predicted ModE elements with scores $>\mu_i - 2\sigma_i$.

In addition to their sequence divergence, the genomes of *E. coli* and *G. sulfurreducens* differ in their average GC content, which may affect the specificity of the binding site matrices. The GC content of the *E. coli* genome is close to 50% (Robison et al., 1998), while that for *G. sulfurreducens* is 61% (Yan et al., 2004). Due to the difference between *E. coli* and *G. sulfurreducens*, investigating candidate sites at a lower cutoff level for ScanACE scores, e.g., $>\mu_i - 2\sigma_i$, may be essential for some searches in order not to miss true positive binding sites. However, a considerable number of false positive hits are found due to both the difference in nucleotide composition of the upstream and the coding regions of the genome, and the sequence divergence of the *E. coli* and *G. sulfurreducens* genomes. Tan et al. (2001) employed the distribution of random sites and the proportion of sites located in the upstream regions in order to select the optimal cutoff values in search for transcription regulatory sites in *Haemophilus influenzae* genome using *E. coli* CRP and FNR binding site matrices. Our results suggest that the required cutoff values in *G. sulfurreducens* may vary substantially among binding sites for different transcription factors, depending

on divergence of regulatory mechanisms and transcription factor specificity between *G. sulfurreducens* and *E. coli*. Experimental verification of predicted elements may help select the optimum cutoff level for individual binding site matrices that would identify a sufficient number of true positive sites while eliminating the false positive ones. Additional criteria such as binding site mutation rate (Brown and Callan, 2004), which has been successfully employed by other authors to provide statistical significance of predicted binding sites in γ -Proteobacteria (Guía et al., 2005), may aid in establishing the cutoff values for binding site predictions in *G. sulfurreducens*.

Due to evolutionary divergence among species lineages, some transcription factors recognize elements that are so distinct among species that they cannot be detected by direct similarity searches. For example, LexA recognizes a palindromic sequence specific to *Geobacter* (Jara et al., 2003). Experimental and computational methods other than direct similarity searches are necessary to identify such unique regulatory elements. Sequence comparison of *G. sulfurreducens* transcription factors to those in *E. coli* and other bacteria may suggest which of their DNA recognition sites are substantially different. For example, sequence comparisons of *G. sulfurreducens* and *E. coli* RpoS and RpoD have shown conservation of domains responsible for promoter recognition and DNA binding and also indicated several functionally important sites that are different (Yan et al., 2006). Comparisons have also been performed for several regulators of response to nitric oxide, including the comparison of the *G. sulfurreducens* HcpR to *E. coli* CRP and FNR (Rodionov et al., 2005). Future studies may identify those transcription factors in which their DNA recognition domains mutated so much that they have lost their ability to recognize DNA sequence elements similar to those of *E. coli*.

The present analysis of transcription factor homology between *E. coli* and *Geobacter* identified those transcription factors that are likely absent from the *G. sulfurreducens* genome. Additional phylogenetic analysis of the subunits of HU and IHF, and of homologs of RpoE in *G. sulfurreducens* and *E. coli* showed the complexity of relationships among transcription factors compared between these species. Future phylogenetic analysis of other gene families of *Geobacter* transcription factors, coupled with functional studies, will provide further detailed insights into their orthology relationships to known transcription factors from other bacteria.

Our search results provided in the online supplement show multiple elements located in proximity to one another. Some of these sites may co-occur due to

cooperative binding of transcription regulators (Barnard et al., 2004), while others may reflect alternative mechanisms of regulation of gene expression under different conditions (Rhodius and LaRossa, 2003). Additionally, some of the predicted elements partially overlap, and they may represent fragments of the same transcription factor binding site.

Ongoing and future experimental studies will verify the accuracy of computational predictions of transcription regulatory elements using similarity searches. Some predictions may be further filtered by incorporating additional sequence information from adjacent regions, for example, about the proximity of the transcription initiation sites, which has been applied by other authors to *E. coli* genome analysis (Thieffry et al., 1998). In addition to 59 transcription factors investigated in this study, many other regulatory proteins exist in *E. coli* (Robison et al., 1998; Thieffry et al., 1998) and in other bacterial species. While the use of footprinted binding site matrices makes it possible to scan a genome for similar binding sites of known transcription factors, other methods are needed to identify binding sites for species-specific transcription factors or for species-specific recognition sites. Such methods include, for example, analyses of co-regulated operons identified from microarray expression data, phylogenetic footprinting, use of energy constraints, and genome-wide searches for overrepresented elements (Vanet et al., 2000; McCue et al., 2001; Li et al., 2002; Djordjevic et al., 2003; Rhodius and LaRossa, 2003; Brown and Callan, 2004).

Our recent analyses of *G. sulfurreducens* gene expression microarray data confirmed a number of regulatory elements reported in the online supplement. Two of the many confirmed sites include Fur sites at positions 1013986–1013999 upstream of operon 432 encoding cystathionine beta-lyases and at positions 2735603–2735623 upstream of transcription unit 1099 encoding a putative oxalate/formate antiporter (J. Krushkal, B. Yan, M. Coppi, L. DiDonato, R. Mahadevan, R. O'Neil, B. Methé, D. Lovley, unpublished data). A number of –35/–10 promoters predicted using sequence similarity searches were validated using microarray analysis of the *rpoS* deletion mutant (Yan et al., 2006). Furthermore, two likely RpoS-regulated promoters upstream of transcription unit 161 encoding an OmpA domain protein and 702 encoding a methylamine utilization protein have been experimentally verified using primer extension analysis (Yan et al., 2006).

Because different regulatory mechanisms may affect gene expression under different biological conditions,

each approach to prediction of regulatory elements (e.g., similarity searches, microarray analyses, and searches for binding sites conserved among species) may identify only a subset of all regulatory elements present in a genome. A combination of theoretical and experimental approaches is necessary in order to investigate the complex interplay of regulatory elements affecting gene expression and to identify those regulatory elements that play a role in particular biological conditions.

Acknowledgements

We thank anonymous reviewers for helpful comments. We are grateful to Dr. R. Adkins (University of Tennessee, Memphis) for helpful suggestions and assistance with manuscript preparation. We also thank Drs. R. Mahadevan (University of Toronto), J. Blanchard (University of Massachusetts, Amherst) and A. Esteve-Núñez (Instituto Nacional de Técnica Aeroespacial, Madrid) for helpful discussions, E. Webb (University of Tennessee, Memphis) for help with manuscript preparation, and P. Brown (University of Massachusetts, Amherst) for assistance with Web page presentation.

This research was supported by the Office of Science (BER), U.S. Department of Energy, Grant No. DE-FC02-02ER63446.

References

- Azam, T.A., Ishihama, A., 1999. Twelve species of the nucleoid-associated protein from *Escherichia coli*: sequence recognition specificity and DNA binding affinity. *J. Biol. Chem.* 274, 33105–33113.
- Barnard, A., Wolfe, A., Busby, S., 2004. Regulation at complex bacterial promoters: how bacteria use different promoter organizations to produce different regulatory outcomes. *Curr. Opin. Microbiol.* 7, 102–108.
- Berg, O.G., von Hippel, P.H., 1988. Selection of DNA binding sites by regulatory proteins. II. The binding specificity of cyclic AMP receptor protein to recognition sites. *J. Mol. Biol.* 200, 709–723.
- Blattner, F.R., Plunkett 3rd., G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B., Shao, Y., 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* 277, 1453–1474.
- Bond, D.R., Lovley, D.R., 2003. Electricity production by *Geobacter sulfurreducens* attached to electrodes. *Appl. Environ. Microbiol.* 69, 1548–1555.
- Brown, C.T., Callan Jr., C.G., 2004. Evolutionary comparisons suggest many novel cAMP response protein binding sites in *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.A.* 101, 2404–2409.
- Caccavo, F.J., Lonergan, D.J., Lovley, D.R., Davis, M., Stolz, J.F., McInerney, M.J., 1994. *Geobacter sulfurreducens* sp. nov., a hydrogen- and acetate-oxidizing dissimilatory metal-reducing microorganism. *Appl. Environ. Microbiol.* 60, 3752–3759.

- Chen, Q.K., Hertz, G.Z., Stormo, G.D., 1995. MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. *Comput. Appl. Biosci.* 11, 563–566.
- Delcher, A.L., Harmon, D., Kasif, S., White, O., Salzberg, S.L., 1999. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* 27, 4636–4641.
- Djordjevic, M., Sengupta, A.M., Shraiman, B.I., 2003. A biophysical approach to transcription factor binding site discovery. *Genome Res.* 13, 2381–2390.
- Eggers, C.H., Caimano, M.J., Radolf, J.D., 2006. Sigma factor selectivity in *Borrelia burgdorferi*: RpoS recognition of the *ospE/ospF/lep* promoters is dependent on the sequence of the -10 region. *Mol. Microbiol.* 59, 1859–1875.
- Esteve-Nunez, A., Nunez, C., Lovley, D.R., 2004. H-034. Regulation of fumarate respiration by CRP-FNR-like proteins in *Geobacter sulfurreducens*. In: 104th General Meeting of the American Society for Microbiology, New Orleans, LA, p. 292.
- Felsenstein, J., 1989. PHYLIP—phylogeny inference package (Version 3.2). *Cladistics* 5, 164–166.
- Fogel, G.B., Weekes, D.G., Varga, G., Dow, E.R., Craven, A.M., Harlow, H.B., Su, E.W., Onyia, J.E., Su, C., 2005. A statistical analysis of the TRANSFAC database. *Biosystems* 81, 137–154.
- Froehlich, B., Husmann, L., Caron, J., Scott, J.R., 1994. Regulation of *rns*, a positive regulatory factor for pili of enterotoxigenic *Escherichia coli*. *J. Bacteriol.* 176, 5385–5392.
- Graur, D., Li, W.-H., 2000. *Fundamentals of Molecular Evolution*, second ed. Sinauer, Sunderland, MA, pp. 200–202.
- Gruber, T.M., Gross, C.A., 2003. Multiple sigma subunits and the partitioning of bacterial transcription space. *Annu. Rev. Microbiol.* 57, 441–466.
- Guía, M.H., Pérez, A.G., Angarica, V.E., Vasconcelos, A.T., Collado-Vides, J., 2005. Complementing computationally predicted regulatory sites in Tractor.DB using a pattern matching approach. *In Silico Biol.* 5, 209–219.
- Hall, B.G., 2004. *Phylogenetic Trees Made Easy: A How-To Manual*, second ed. Sinauer, Sunderland MA, MA, pp. 55–56.
- Hengge-Aronis, R., 2000. The general stress response in *Escherichia coli*. In: Storz, G., Hengge-Aronis, R. (Eds.), *Bacterial Stress Responses*. ASM Press, Washington, DC, pp. 161–178.
- Hengge-Aronis, R., 2002. Stationary phase gene regulation: what makes an *Escherichia coli* promoter σ^S -selective? *Curr. Opin. Microbiol.* 5, 591–595.
- Holmes, D.E., Finneran, K.T., O'Neil, R.A., Lovley, D.R., 2002. Enrichment of members of the family *Geobacteraceae* associated with stimulation of dissimilatory metal reduction in uranium-contaminated aquifer sediments. *Appl. Environ. Microbiol.* 68, 2300–2306.
- Homuth, G., Masuda, S., Mogk, A., Kobayashi, Y., Schumann, W., 1997. The *dnaK* operon of *Bacillus subtilis* is heptacistronic. *J. Bacteriol.* 179, 1153–1164.
- Huerta, A.M., Collado-Vides, J., 2003. Sigma70 promoters in *Escherichia coli*: specific transcription in dense regions of overlapping promoter-like signals. *J. Mol. Biol.* 333, 261–278.
- Jara, M., Núñez, C., Campoy, S., Fernandez de Henestrosa, A.R., Lovley, D.R., Barbé, J., 2003. *Geobacter sulfurreducens* has two autoregulated *lexA* genes whose products do not bind the *recA* promoter: differing responses of *lexA* and *recA* to DNA damage. *J. Bacteriol.* 185, 2493–2502.
- Karp, P.D., Riley, M., Saier, M., Paulsen, I.T., Collado-Vides, J., Paley, S.M., Pellegrini-Toole, A., Bonavides, C., Gama-Castro, S., 2002. The EcoCyc database. *Nucleic Acids Res.* 30, 56–58.
- Kormanec, J., Farkašovsky, M., 1993. Differential expression of principal sigma factor homologues of *Streptomyces aureofaciens* correlates with the developmental stage. *Nucleic Acids Res.* 21, 3647–3652.
- Kumar, S., Tamura, K., Nei, M., 2004. MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief. Bioinform.* 5, 150–163.
- Lacour, S., Landini, P., 2004. σ^S -Dependent gene expression at the onset of stationary phase in *Escherichia coli*: function of σ^S -dependent genes and identification of their promoter sequences. *J. Bacteriol.* 186, 7186–7195.
- Lacour, S., Kolb, A., Landini, P., 2003. Nucleotides from -16 to -12 determine specific promoter recognition by bacterial σ^S -RNA polymerase. *J. Biol. Chem.* 278, 37160–37168.
- Laine, B., Kmiecik, D., Sautiere, P., Biserte, G., Cohen-Solal, M., 1980. Complete amino-acid sequences of DNA-binding proteins HU-1 and HU-2 from *Escherichia coli*. *Eur. J. Biochem.* 103 (3), 447–461.
- Leang, C., Lovley, D., 2005. Differential transcriptional regulation and function of two highly similar genes, *omcB* and *omcC*, in a 10-kb chromosomal duplication in *Geobacter sulfurreducens*. *Microbiology* 151, 1761–1767.
- LeBlanc, H., Lang, A.S., Beatty, J.T., 1999. Transcript cleavage, attenuation, and an internal promoter in the *Rhodobacter capsulatus puc* operon. *J. Bacteriol.* 181, 4955–4960.
- Lee, S.J., Gralla, J.D., 2001. Sigma 38 (rpoS) RNA polymerase promoter engagement via -10 region nucleotides. *J. Biol. Chem.* 276, 30064–30071.
- Li, H., Rhodius, V., Gross, C., Siggia, E.D., 2002. Identification of the binding sites of regulatory proteins in bacterial genomes. *Proc. Natl. Acad. Sci. U.S.A.* 99, 11772–11777.
- Liu, X., De Wulf, P., 2004. Probing the ArcA-P modulon of *Escherichia coli* by whole genome transcriptional analysis and sequence recognition profiling. *J. Biol. Chem.* 279, 12588–12597.
- Lonergan, D.J., Jenter, H.L., Coates, J.D., Phillips, E.J., Schmidt, T.M., Lovley, D.R., 1996. Phylogenetic analysis of dissimilatory Fe(III)-reducing bacteria. *J. Bacteriol.* 178, 2402–2408.
- Lovley, D.R., 1997. Microbial Fe(III) reduction in subsurface environments. *FEMS Microbiol. Rev.* 20, 305–313.
- Lovley, D.R., 2002. Analysis of the genetic potential and gene expression of microbial communities involved in the in situ bioremediation of uranium and harvesting electrical energy from organic matter. *OMICS* 6, 331–339.
- Lovley, D.R., 2003. Cleaning up with genomics: applying molecular biology to bioremediation. *Nat. Rev. Microbiol.* 1, 35–44.
- Maeda, H., Fujita, N., Ishihama, A., 2000. Competition among seven *Escherichia coli* σ subunits: relative binding affinities to the core RNA polymerase. *Nucleic Acids Res.* 28, 3497–3503.
- Martinez-Bueno, M., Molina-Henares, A.J., Pareja, E., Ramos, J.L., Tobes, R., 2004. BacTregulators: a database of transcriptional regulators in bacteria and archaea. *Bioinformatics* 20, 2787–2791.
- McCue, L.A., Thompson, W., Carmack, C.S., Ryan, M.P., Liu, J.S., Derbyshire, V., Lawrence, C.E., 2001. Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res.* 29, 774–782.
- McCue, L.A., Thompson, W., Carmack, C.S., Lawrence, C.E., 2002. Factors influencing the identification of transcription factor binding sites by cross-species comparison. *Genome Res.* 12, 1523–1532.
- McNicholas, P.M., Rech, S.A., Gunsalus, R.P., 1997. Characterization of the ModE DNA-binding sites in the control regions of *modABCD* and *moaABCDE* of *Escherichia coli*. *Mol. Microbiol.* 23, 515–524.

- Méthé, B.A., Nelson, K.E., Eisen, J.A., Paulsen, I.T., Nelson, W., Heidelberg, J.F., Wu, D., Wu, M., Ward, N., Beanan, M.J., Dodson, R.J., Madupu, R., Brinkac, L.M., Daugherty, S.C., DeBoy, R.T., Durkin, A.S., Gwinn, M., Kolonay, J.F., Sullivan, S.A., Haft, D.H., Selengut, J., Davidsen, T.M., Zafar, N., White, O., Tran, B., Romero, C., Forberger, H.A., Weidman, J., Khouri, H., Feldblyum, T.V., Utterback, T.R., Van Aken, S.E., Lovley, D.R., Fraser, C.M., 2003. The genome of *Geobacter sulfurreducens*: insights into metal reduction in subsurface environments. *Science* 302, 1967–1969.
- Núñez, C., Adams, L., Childers, S., Lovley, D.R., 2004. The RpoS sigma factor in the dissimilatory Fe(III)-reducing bacterium *Geobacter sulfurreducens*. *J. Bacteriol.* 186, 5543–5546.
- O'Neil, R.A., Coppi, M.V., Nevin, K.P., Woodard, J., Methe, B.A., Webster, J., Krushkal, J., Yan, B., Lovley, D.R., 2004. H-061. Investigation of the Fur regulon of *Geobacter sulfurreducens*. In: 104th General Meeting of the American Society for Microbiology, New Orleans, LA, p. 295.
- Panina, E.M., Mironov, A.A., Gelfand, M.S., 2001. Comparative analysis of FUR regulons in gamma-proteobacteria. *Nucleic Acids Res.* 29, 5195–5206.
- Plumbridge, J., 2001. DNA binding sites for the Mlc and NagC proteins: regulation of *nagE*, encoding the *N*-acetylglucosamine-specific transporter in *Escherichia coli*. *Nucleic Acids Res.* 29, 506–514.
- Prüß, B.M., Campbell, J.W., Van Dyk, T.K., Zhu, C., Kogan, Y., Matsumura, P., 2003. FlhD/FlhC is a regulator of anaerobic respiration and the Entner-Doudoroff pathway through induction of the methyl-accepting chemotaxis protein Aer. *J. Bacteriol.* 185, 534–543.
- Remm, M., Storm, C.E., Sonnhammer, E.L., 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* 314, 1041–1052.
- Rhodijs, V.A., LaRossa, R.A., 2003. Uses and pitfalls of microarrays for studying transcriptional regulation. *Curr. Opin. Microbiol.* 6, 114–119, doi:10.1016/S1369-5274(1003)00034-00031.
- Rice, P., Longden, I., Bleasby, A., 2000. EMBOSS: the European molecular biology open software suite. *Trends Genet.* 16, 276–277.
- Robison, K., McGuire, A.M., Church, G.M., 1998. A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K-12 genome. *J. Mol. Biol.* 284, 241–254.
- Rodionov, D.A., Dubchak, I., Arkin, A., Alm, E., Gelfand, M.S., 2004. Reconstruction of regulatory and metabolic pathways in metal-reducing δ -proteobacteria. *Genome Biol.* 5, R90, doi:10.1186/gb-2004-1185-11f11-r1190.
- Rodionov, D.A., Dubchak, I.L., Arkin, A.P., Alm, E.J., Gelfand, M.S., 2005. Dissimilatory metabolism of nitrogen oxides in bacteria: comparative reconstruction of transcriptional networks. *PLoS Comput. Biol.* 1, e55.
- Roth, F.P., Hughes, J.D., Estep, P.W., Church, G.M., 1998. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.* 16, 939–945.
- 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., Collado-Vides, J., 2004. RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. *Nucleic Acids Res.* 32 (Database issue), D303–D306.
- Studholme, D.J., Pau, R.N., 2003. A DNA element recognised by the molybdenum-responsive transcription factor ModE is conserved in Proteobacteria, green sulphur bacteria and Archaea. *BMC Microbiol.* 3, 24.
- Tan, K., Moreno-Hagelsieb, G., Collado-Vides, J., Stormo, G.D., 2001. A comparative genomics approach to prediction of new members of regulons. *Genome Res.* 11, 566–584.
- Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D., Koonin, E.V., 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* 29, 22–28.
- Thieffry, D., Salgado, H., Huerta, A.M., Collado-Vides, J., 1998. Prediction of transcriptional regulatory sites in the complete genome sequence of *Escherichia coli* K-12. *Bioinformatics* 14, 391–400.
- Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G., 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res. (Online)* 25, 4876–4882.
- Vaillancourt, K., Moineau, S., Frenette, M., Lessard, C., Vadeboncoeur, C., 2002. Galactose and lactose genes from the galactose-positive bacterium *Streptococcus salivarius* and the phylogenetically related galactose-negative bacterium *Streptococcus thermophilus*: organization, sequence, transcription, and activity of the *gal* gene products. *J. Bacteriol.* 184, 193–785.
- Vanet, A., Marsan, L., Labigne, A., Sagot, M.-F., 2000. Inferring regulatory elements from a whole genome. An analysis of *Helicobacter pylori* σ^{80} family of promoter signals. *J. Mol. Biol.* 297, 335–353, doi:10.1006/jmbi.2000.3576.
- Weber, H., Polen, T., Heuveling, J., Wendisch, V.F., Hengge, R., 2005. Genome-wide analysis of the general stress response network in *Escherichia coli*: σ^S -dependent genes, promoters, and sigma factor selectivity. *J. Bacteriol.* 187, 1591–1603.
- Weglenska, A., Jacob, B., Sirko, A., 1996. Transcriptional pattern of *Escherichia coli* *ihfB* (*himD*) gene expression. *Gene* 181, 85–88.
- Yan, B., Méthé, B.A., Lovley, D.R., Krushkal, J., 2004. Computational prediction of conserved operons and phylogenetic footprinting of transcription regulatory elements in the metal-reducing bacterial family *Geobacteraceae*. *J. Theor. Biol.* 230, 133–144, doi:10.1016/j.jtbi.2004.1004.1022.
- Yan, B., Núñez, C., Ueki, T., Esteve-Núñez, A., Puljic, M., Adkins, R.M., Méthé, B.A., Lovley, D.R., Krushkal, J., 2006. Computational prediction of RpoS and RpoD regulatory sites in *Geobacter sulfurreducens* using sequence and gene expression information. *Gene* 384, 73–95, doi:10.1016/j.gene.2006.1006.1025.
- Zheng, J., Wu, J., Sun, Z., 2003. An approach to identify over-represented cis-elements in related sequences. *Nucleic Acids Res.* 31, 1995–2005.