Nucleic Acids Research Advance Access published October 26, 2006

Nucleic Acids Research, 2006, Vol. 00, No. 00 **1–14**
doi:10.1093/nar/gkl734

Non-canonical CRP sites control competence regulons in *Escherichia coli* and many other γ-proteobacteria

Andrew D. S. Cameron[1] and Rosemary J. Redfield[2,*]

[1]Department of Microbiology and Immunology and [2]Department of Zoology, University of British Columbia, Vancouver, BC, Canada

Received August 14, 2006; Revised September 20, 2006; Accepted September 21, 2006

ABSTRACT

***Escherichia coli*'s cAMP receptor protein (CRP), the archetypal bacterial transcription factor, regulates over a hundred promoters by binding 22 bp symmetrical sites with the consensus core half-site TGTGA. However, *Haemophilus influenzae* has two types of CRP sites, one like *E.coli*'s and one with the core sequence TG<u>C</u>GA that regulates genes required for DNA uptake (natural competence). Only the latter 'CRP-S' sites require both CRP and the coregulator Sxy for activation. To our knowledge, the TG<u>T</u>GA and TG<u>C</u>GA motifs are the first example of one transcription factor having two distinct binding-site motifs. Here we show that CRP-S promoters are widespread in the γ-proteobacteria and demonstrate their Sxy-dependence in *E.coli*. Orthologs of most *H.influenzae* CRP-S-regulated genes are ubiquitous in the five best-studied γ-proteobacteria families, *Enterobacteriaceae*, *Pasteurellaceae*, *Pseudomonadaceae*, *Vibrionaceae* and *Xanthomonadaceae*. Phylogenetic footprinting identified CRP-S sites in the promoter regions of the *Enterobacteriaceae*, *Pasteurellaceae* and *Vibrionaceae* orthologs, and canonical CRP sites in orthologs of genes known to be Sxy-independent in *H.influenzae*. Bandshift experiments confirmed that *E.coli* CRP-S sequences are low affinity binding sites for CRP, and mRNA analysis showed that they require CRP, cAMP (CRP's allosteric effector) and Sxy for gene induction. This work suggests not only that the γ-proteobacteria share a common DNA uptake mechanism, but also that, in the three best studied families, their competence regulons share both CRP-S specificity and Sxy dependence.**

INTRODUCTION

The *Escherichia coli* cAMP receptor protein (CRP), also called the catabolite activator protein (CAP), was the first transcription factor to be purified and the first to have its structure solved (1,2). The protein's N-terminal sensory domain binds its allosteric effector cyclic AMP (cAMP) with high affinity, resulting in a conformational change that exposes a C-terminal helix–turn–helix DNA-binding domain. Adenylate cyclase raises intracellular levels of cAMP sufficiently to trigger CRP-DNA binding when the flow of preferred (PTS-transported) sugars across the cell membrane slows or stops, usually because of depletion of these sugars in the cell's environment. Once bound to DNA, CRP makes protein–protein contacts with RNA polymerase and recruits it to promoters to initiate transcription. In rare cases CRP acts as a repressor by overlapping polymerase-binding sites (3). Over 100 CRP-regulated promoters have been identified experimentally (listed at RegulonDB, http://regulondb.ccg.unam.mx:80/index.html) and over 400 sites have been predicted computationally (4) (listed at TractorDB, http://www.tractor.lncc.br/), making CRP the global regulator of the cell's response to carbon and energy shortage.

E.coli CRP binds as a homodimer, specifically to symmetrical 22 bp DNA sites with the consensus half site 5′-$A_1A_2A_3\mathbf{T_4G_5T_6G_7A_8}T_9C_{10}T_{11}$. The protein makes direct contact with base pairs $G{:}C_5$, $G{:}C_7$ and $A{:}T_8$ in the highly conserved core motif $T_4G_5T_6G_7A_8$, and binding induces a localized kink of 43° between positions 6 and 7, wrapping the DNA around CRP and strengthening the association (5,6). Though base pair $T{:}A_6$ is not directly contacted by CRP, it is recognized indirectly because kink formation strongly favours $T{:}A_6$ over other base pairs (5–7). For example, replacement of $T{:}A_6$ in a consensus CRP site with $C{:}G_6$ causes an 80-fold reduction in CRP affinity by increasing the free energy required to bend the DNA (6).

In vitro, transcription stimulation by *E.coli* CRP requires no other protein factors (8). *In vivo*, however, CRP-regulated promoters are typically coregulated by one or more additional

*To whom correspondence should be addressed at Life Sciences Centre (Zoology), 2350 Health Sciences Mall, University of British Columbia, Vancouver, BC, Canada V6T 1Z3. Tel: +604 822 3744; Fax: +604 827 4135; Email: redfield@zoology.ubc.ca

factors binding to DNA sites adjacent to CRP. The classic example is the *lacZYA* promoter, which contains binding sites for both CRP and the LacI repressor. Although, CRP binds to this promoter during sugar starvation, no transcription occurs unless the LacI repressor binds lactose and releases the DNA. Many other interactions have been characterized (9) (see RegulonDB for a list of CRP's coregulators). Some coregulators act independently of CRP; others affect CRP binding either by modifying DNA conformation or by increasing the local CRP concentration through protein–protein contacts. This complex interplay between multiple regulators at any given promoter may explain why Zheng and coworkers found that the degree of promoter dependence on CRP was not correlated with the quality of the CRP-binding site (3).

CRP-DNA affinity increases with increasing similarity of a DNA site to the CRP consensus, but CRP's affinity for a site matching the consensus is too strong to be biologically useful (10). This may explain why none of the 182 experimentally determined *E.coli* CRP sites listed in RegulonDB exactly match the 22 nt consensus and all but nine sites are mismatched at one or more positions of the 10 nt core. The degree of similarity to the consensus has been proposed to generate an adaptive hierarchy allowing genes with better sites to be preferentially activated at low cAMP concentrations (11,12).

Despite the extensive variation among CRP sites, no significance has been attached to which positions vary. However, this model is changing with the new understanding of CRP-binding site specificity emerging from studies in the naturally competent bacterium *Haemophilus influenzae*. Transcriptome analysis of competence-inducing conditions in *H.influenzae* revealed that, in addition to the expected suite of CRP-promoters with typical CRP sites, unusual CRP-binding sites regulate genes required for DNA uptake (13). The CRP sites in these 13 competence-induced promoters are described by an alternative motif, 5′-$T_1T_2T_3\mathbf{T_4G_5C_6G_7}$-$\mathbf{A_8}T_9C_{10}T_{11}$ (note $\mathbf{C_6}$ rather than $\mathbf{T_6}$), and absolutely require a second protein, Sxy (also called TfoX), for induction. Because Sxy lacks recognizable DNA-binding domains, and Sxy-dependent promoters contain no other sequence motifs, Sxy is not thought to act by binding a specific DNA sequence. Instead, the presence of C rather than T at position 6 of the CRP half-site appears to make Sxy essential for CRP-DNA binding and transcription activation (13,14). Consistent with this requirement, conditions that induce competence increase *sxy* expression, and *sxy* over-expression leads to strong induction of the competence genes (13,15). Because these competence-specific CRP-binding sites were originally identified only as consensus sequences in *H.influenzae* competence gene promoters, they were called competence regulatory elements (CREs). Here we introduce the terms CRP-N and CRP-S to distinguish between canonical (Sxy-independent) and Sxy-dependent CRP sites.

Natural competence is known in only a few γ-proteobacteria [*Vibrio cholerae*, five *Pasteurellaceae* species and three species of *Pseudomonas* (16–18)], and our understanding of its genetics and molecular mechanisms comes almost exclusively from studies of *H.influenzae*, where genetic analysis has identified more than 20 genes required for DNA-binding, transport and recombination [e.g. (19,20), summarized in (13)]. Here we report that competence is likely to be ubiquitous in the γ-proteobacteria, as most of the genes essential for competence and transformation in *H.influenzae* are found in the five best-studied γ-proteobacteria families (*Enterobacteriaceae*, *Pasteurellaceae*, *Pseudomonadaceae*, *Vibrionaceae* and *Xanthomonadaceae*). In three of these families (*Enterobacteriaceae*, *Pasteurellaceae* and *Vibrionaceae*), many of these genes have promoter sites matching the *H.influenzae* CRP-S motif. In *E.coli,* we demonstrate experimentally that these CRP-S promoters, like their *H.influenzae* counterparts, require both CRP and Sxy for transcription.

MATERIALS AND METHODS

Genome sequence analysis

Sequences from the complete and annotated genomes of *E.coli K12*-MG1655, *H.influenzae KW20 Rd, Haemophilus ducreyi 35000HP, Mannheimia succiniciproducens MBEL55E, Pasteurella multocida Pm70, Pseudomonas aeruginosa PAO1, Pseudomonas fluorescens Pf-5, Salmonella typhimurium LT2 SGSC1412, V.cholerae El Tor N16961, Vibrio parahaemolyticus RIMD 2210633, Vibrio vulnificus YJ016, Yersinia pestis KIM, Xanthomonas campestris pv. campestris ATCC33913, Xylella fastidiosa 9a5c* were retrieved from The Institute for Genomic Research (TIGR, http://www.tigr.org). The complete *Haemophilus somnus 129-PT* and unfinished *H.somnus* 2336 genomes were retrieved from http://www.jgi.doe.gov and http://www.ncbi.nlm.nih.gov, respectively. The unfinished genomes of *Actinobacillus actinomycetemcomitans HK1651* and *Actinobacillus pleuropneumoniae* serovar *1 strain 4074* were retrieved from NCBI (http://www.ncbi.nlm.nih.gov). Sequence from the unfinished *Mannheimia haemolytica PHL213* genome was obtained from the Baylor College of Medicine Human Genome Sequencing Center (http://www.hgsc.bcm.tmc.edu). Some searches included five additional *Pseudomonadaceae* genomes (*Pseudomons syringae, Pseudomons fluorescens PfO-1, Pseudomons putida KT2440, P.syringae phaseolicola 1448A*, and *P.syringae pv B728a*) and five additional *Xanthomonadaceae* genomes (*Xanthomonas citri, Xanthomonas campestris 8004, X.campestris vesicatoria 85-10, Xanthomonas fastidiosa Temecula1, Xanthomonas oryzae KACC10331*).

Completed genomes were searched using BLASTP and incomplete genomes were searched using TBLASTN. The *M.haemolytica* genome was searched using the BLAST server at Baylor College of Medicine; all other searches were conducted using the NCBI and TIGR web servers. For unfinished genomes, open reading frames were visualized using Sequence Analysis (http://informagen.com/SA/). Genes were considered orthologous if they were the top hit in reciprocal BLAST searches and if the alignment included at least 75% of the shorter gene. All homologs of *H.influenzae* CRP-S regulon genes identified in this study fit this definition, except some of those in the *comA-E* and the *pulG*-HI0941 operons. The *comA-E* operon has been previously shown to be conserved in γ-proteobacteria (21). For several homologs of *H.influenzae* CRP-N-regulated genes, duplication events have generated paralogs in some species, thus we analysed

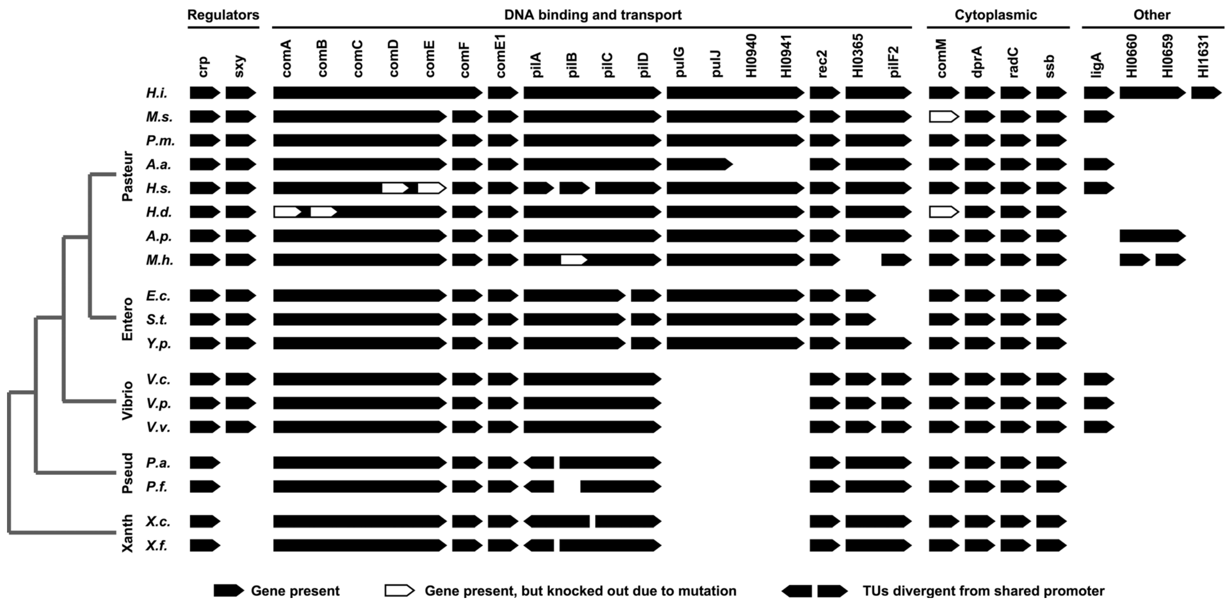

Figure 1. Orthologs of *H.influenzae* CRP-S-regulated genes in other γ-proteobacteria. Solid lines depict transcriptional units (gene lengths not to scale). Cladogram adapted from Lerat *et al*. (33). Abbreviations: Pasteur, *Pasteurellaceae*; Entero, *Enterobacteriaceae*; Vibrio, *Vibrionaceae*; Pseud, *Pseudomonadaceae*; Xanth, *Xanthomonadaceae*; H.i., *H.influenzae*; M.s., *M.succiniciproducens*; P.m., *P.multocida*; A.a., *A.actinomycetemcomitans*; H.s., *H.somnus*; A.p., *A.pleuropneumoniae*; M.h., *M.haemolytica*; E.c. *E.coli*; S.t., *S.typhimurium*; Y.p., *Y.pestis*; V.c., *V.cholerae*; V.p., *V.parahaemolyticus*; V.v., *V.vulnificus*; P.a., *P.aeruginosa*; P.f., *P.fluorescens*; X.c., *X.campestris*; X.f., *X.fastidiosa*.

Table 1. Details of phylogenetic footprinting.

Family	Genomes	Orthologs of *H.influenzae* CRP-S genes Promoters searched	Motifs found	Sites found	Orthologs of *H.influenzae* CRP-N genes Promoters searched	Motifs found	Sites found
Pasteur	8	91	1	87	109[a]	1	116
Entero	3	33	1	38	90	1	57
Vibrio	3	33	0	0	71	2	i. 49 ii. 27
		15[b]	1	24			
Pseudo	7	63	0	0	119	0	0
Xantho	7	68	0	0	77	0	0

[a]Includes only *H.influenzae*, *M.succiniciproducens*, *P.multocida* and *H.ducreyi* promoters.
[b]Results of an alternate search strategy employed for *Vibrionaceae* (explained in Results).

all paralog promoters. For the *Pseudomonadaceae* and *Xanthomonadaceae* species not listed in Figure 1, gene orthologs were identified using RSATools 'ortholog search' (http://rsat.ulb.ac.be/rsat/) (22).

Promoter analysis: identifying transcription factor binding sites

Promoter regions were defined as the sequence between −300 bp and the start codon of the first gene in a transcription unit. The *H.influenzae comA-E* operon CRP-S site overlaps an upstream ORF, so we allowed overlap with upstream ORFs in all searches to avoid missing transcription factor binding sites. In cases where gene order within transcriptional units differs between lineages, we analysed only the promoter regions of predicted transcription units, and not the DNA immediately upstream of orthologs.

CONSENSUS (23) and Gibbs motif sampler (24) were run using RSATools. BioProspector (25) was run at http://ai.stanford.edu/~xsliu/BioProspector/. Because motif discovery algorithms have poor accuracy when searching for motifs shorter than 10 bp (26), we tested the following motif widths: 10, 11, 12, 13, 14, 16, 18, 20 bp for BioProspector, plus 22 bp for CONSENSUS and Gibbs. Sites identified by all three programs as matching a significant motif in all width categories were included in Table 1. The average *E.coli* transcription factor binding site motif length is 21 (26), and statistical significance is greater for longer motifs due to increased information content; thus special consideration was given to sites identified only in search widths greater than 16 bp if they were identified in all 18 to 22 bp searches.

Parameters were set to allow for promoters with multiple or no sites. BioProspector was set to search for either one block motifs, or two-block palindromes with a gap of 0 to 6 bases between blocks; background models were set as '*E.coli* intergenic' for searching *Enterobacteriaceae* and '*V.cholerae* intergenic' for searching *Vibrionaceae*, while background was modeled from the promoters being searched for the other three families. Searching the reverse DNA strand or for symmetrical motifs with CONSENSUS and Gibbs did not identify any additional high-confidence sites.

To score putative CRP sites in the *Pasteurellaceae*, three weight matrices were generated as described previously (13,14). I_{seq} scores were calculated using PATSER, available at RSATools.

E.coli strains

The pASKA*sxy* clone (JW0942, CmR), and knockouts *crp*::KanR (JWK5702) and *cyaA*::KanR (JWK3778) were acquired from the GenoBase ASKA/GFP(-) and KO collections, respectively (27,28), and cultured on Luria–Bertani (LB) medium (30 μg/ml chloramphenicol or 10 μg/ml kanamycin). Knockout strains were made chemically competent with RbCl and transformed with pASKA*sxy* as described previously (29).

Protein purification and bandshifts

E.coli CRP was purified from a strain constructed by Peekhaus and Conway (30) in which the *crp* coding sequence is cloned under *lac* promoter control in the His-tag vector pQE30 (Qiagen). Cells were grown in LB (25 μg/ml kanamycin and 100 μg/ml ampicillin) and *crp* expression was induced at OD_{600} 0.6 with 1 mM isopropyl-β-D-thiogalactopyranoside (IPTG). Cells were harvested after 4.5 h by centrifugation and the pellet was frozen overnight at −20°. Native CRP was purified as follows: the pellet was resuspended in lysis buffer (50 mM sodium phosphate, 300 mM sodium chloride and 10 mM imidazole), then treated with 1 mg/ml lysozyme for 30 min at 24° followed by sonication on ice. Insoluble material was removed by centrifugation at 10 000 *g* for 25 min and the supernatant was then incubated with nickel-nitriloacetic acid agarose beads for 1 h at 4° with gentle rocking. The agarose beads were loaded in a column and washed twice with four column volumes of wash buffer (50 mM sodium phosphate, 300 mM sodium chloride and 20 mM imidazole), and protein was collected in elution buffer (50 mM sodium phosphate, 300 mM sodium chloride and 250 mM imidazole). Purified protein was desalted with Nanosep 3K Omega membranes (Pall), then resuspended in storage buffer (20% glycerol, 40 mM Tris and 200 mM potassium chloride) and stored at −80°. CRP purity was assessed on Coomassie stained SDS–PAGE gels.

PCR was used to amplify DNA fragments containing the *ppdD, yrfD* and *lacZ* CRP sites as well as part of the coding region from *hofB*. The following primers were used for PCR: ppdDF 5′-CGTTTTCGCTAATAGTTGACAG, ppdDR 5′-AGATTCCGAGGTTTTTTATTTC, yrfDF 5′-CGCTGTAA-ATCTGCATCGGA, yrfDR 5′-CAGTCTGTTGCATTCTGC-TGGG, lacZF 5′-GCACGACAGGTTTCCCGACT, lacZR 5′-CACAATTCCACACAACATAC, hofBF 5′-GCCTACCG-CATCCGCTT, hofBR 5′-CCAGGTTTCCCAGCACTTTTA-AT. Amplicons were purified using PAGE. Bands were then excised and DNA was eluted from macerated gel overnight in TE at 37°, ethanol precipitated and resuspended in 10 mM Tris. DNA was end-labeled with T4 polynucleotide kinase using a 10-fold molar excess of [γ-^{32}P]ATP, and unincorporated label was removed with a PCR cleanup spin column (Sigma).

CRP-DNA binding reactions (10 μl) contained 100 nM CRP, 10 mM Tris (pH 8.0), 50 mM KCl, 5% (v/v) glycerol, 250 μg/ml BSA, 100 μM cAMP, 1 mM DTT, 40 μg/μl poly(dI–dC) DNA and 1×10^6 c.p.m. labelled bait DNA. Reactions were incubated at room temperature for 10 min before being loaded onto a prerun polyacrylamide gel [30:1 acrylamide–bisacrylamide; 0.2× TBE (89 mM Tris, 89 mM borate and 2 mM EDTA (pH 8.3)], 2% glycerol, and 200 μM cAMP; running buffer 0.2× TBE and 200 μM cAMP. After electrophoresis for 2.5 h at 100 V, the gel was dried and exposed for 2 h to a phosphor screen. Bands were visualized using a STORM 860 scanner.

Quantitative PCR

Total RNA was isolated from cultures using RNeasy Mini Kits (QIAGEN) and purity and quality assessed by electrophoresis in 1% agarose (1× TAE). RNA was then DNase treated twice with a DNA Free kit (AMBION) and cDNA templates were synthesized using the iScript cDNA synthesis kit (BioRad). PCR primers: *ppdD* primers same as *hofB* primers above, yrfDF 5′-TGGCTGTCAGGGACGATG, yrfDR 5′-ACTGAGTGAGTCTTCGCTGTAATCG, sbmCF 5′-GACGGTGCCGGGTTACTTT, sbmCR 5′-GCATACTG-ACCACCTGTAATTTCTG, mglBF 5′-GTCCAGCATTCC-GGTGTTTGG, mglBR 5′-CGCCTGGTTGTTAGCATCGT. Reactions were carried out in duplicate with each primer set on an ABI 7000 Sequence Detection System (Applied Biosystems) using iTaq SYBR Green Supermix (BioRad). 23S rRNA was used as an internal standard for each RNA prep, with cDNA templates diluted 1/1000 and 1/10 000; 23SF 5′-GCTGATACCGCCCAAGAGTT, 23SR 5′-CAGG-ATGTGATGAGCCGAC. Standard curves were generated with five serial tenfold dilutions of DH5α chromosomal DNA.

Phylogenetic analysis

Amino acid sequences were aligned using CLUSTALX, and these alignments were used to align nucleic acid sequences as codons using Codon Align (31). Phylogenies were estimated using the PHYLIP software package (32). The trees presented in Figure 8 are consensus trees from 100 datasets generated with SeqBoot. Maximum likelihood trees were constructed using dnaML, and parsimony trees were constructed using dnaPars; both programs generated congruent consensus trees (produced with Consense).

RESULTS

The discovery that *H.influenzae* has two kinds of CRP sites with distinct regulatory functions immediately raised the question of whether this dichotomy occurs in other species. This issue is especially pertinent for *E.coli*, where CRP has been thoroughly studied and is thought to be very well characterized. To address this we first identified orthologs of *H.influenzae* CRP-S genes in other genomes and examined their promoter regions for sequence motifs.

Orthologs of *H.influenzae* competence regulon genes in γ-proteobacteria

We have previously reported that all sequenced *Pasteurellaceae* genomes have the 17 genes required for DNA binding and uptake in *H.influenzae* (16). Here we extend this to all 26 of the genes in *H.influenzae*'s CRP-S regulon and to members of four other γ–proteobacteria families: the

Enterobacteriaceae, Pseudomonadaceae, Vibrionaceae and *Xanthomonadaceae*. We have excluded other γ-proteobacterial families from our analysis because they have not been as well studied and lack multiple genome sequences. The five families analysed here have well resolved phylogenies (see tree on the left side of Figure 1) and are used routinely to represent the diversity of γ-proteobacteria (33–35).

Figure 1 shows the results of our expanded search. Orthologs of *crp* are present in all genomes. The competence-specific regulator *sxy* has orthologs in the *Enterobacteriaceae, Pasteurellaceae* and *Vibrionaceae*; in the latter a gene duplication event has generated *sxy* paralogs. In addition, weak matches to the Sxy N- and C-terminal domains (BLAST *E*-values > 0.01) are scattered throughout the eubacteria, suggesting that these domains represent functionally independent modules.

All five families have orthologs of all '*com*' genes, *pilA-D, rec2, dprA* (*smf*), *radC*, HI0365 and *ssb*, although individual genes are missing from some species. *P.fluorescens* lacks *pilB, E.coli* and *S.typhimurium* lack *pilF2*, and *A.actinomycetemcomitans* lacks HI0940 and HI0941. The incomplete *M.haemolytica* genome is missing sequence upstream of *pilF2*, which may explain why no HI0365 ortholog was found. Other genes have a more sporadic distribution. *ligA*, HI0659 and HI0660 occur in only a few genomes, while HI1631 is unique to *H.influenzae*. Although BLAST searching did not detect any *Enterobacteriaceae* homologs of *H.influenzae pulG*-HI0941 genes, in both *Pasteurellaceae* and *Enterobacteriaceae* four similar-sized genes annotated only as 'prepilin peptidase dependent proteins' are adjacent to the highly conserved *recC*. Thus, we consider these *Enterobacteriaceae* genes to be orthologous to *H.influenzae pulG*-HI0941.

Most but not all of these genes are known to have roles in DNA uptake and transformation in *H.influenzae* (13), and their distribution indicates that they were present in the common ancestor of the γ-proteobacteria. Preservation of these genes over hundreds of millions of years suggests that natural competence may be much more common than previously suspected.

Sequence motifs in competence gene promoters

The continuous arrows in Figure 1 depict predicted transcriptional units; the conservation of these operons suggests that selection on functional interactions between gene products has preserved their common regulation (36). We used cross-species sequence comparisons (also called phylogenetic footprinting) to identify conserved transcription factor binding sites in these promoters. This method is based on the premise that natural selection will have conserved the transcription factor binding sites in promoter regions that have elsewhere accumulated neutral mutations, so that finding shared motifs in promoters of orthologous genes is evidence of a conserved regulatory mechanism.

To avoid biasing the results we did not search for CRP-site motifs, but instead used an unbiased search to find any motifs shared between the upstream 'promoter' regions of the transcriptional units in Figure 1 (promoter regions are defined in Materials and Methods). Promoter regions were pooled within each family and were searched using three popular motif discovery programs: CONSENSUS (23), Gibbs motif sampler (24), and BioProspector (25). All three programs are designed to detect patterns ('motifs') in unaligned DNA. Unlike pairwise and multiple alignment algorithms, motif discovery programs can exclude sequence that does not match a motif while also being able to find multiple repeats of a motif in a sequence. CONSENSUS generates weight matrices and calculates a log-likelihood ratio ('information content') to identify related sequences. Gibbs motif sampler iteratively samples motif models and scores individual sites against the models. BioProspector is a variant of the Gibbs sampling algorithm that integrates relationships between adjacent nucleotides. Motif discovery programs often identify false-positive sites; our use of three different algorithms provides cross-validation and greatly reduces the potential for false-positives (26). Consequently we placed high confidence in sites identified by all three programs. Table 1 shows the number of promoters searched within each bacterial family, as well as the outcome of the phylogenetic footprinting analysis. (Search parameters are described in Materials and Methods.) These analyses generated long lists, which are provided as Supplementary Data; below we present only sequence logo versions of the shared motifs.

CRP-S and CRP-N sites in the *Pasteurellaceae*

Phylogenetic footprint analysis of the 91 Pasteurellaceae promoters in Figure 1 identified a single motif shared by 87 promoters; each of which had a single site. Because the *M.haemolytica* genome sequence is incomplete, promoter sequences could not be associated with *comE1, pilF2* or *comM*. A sequence logo summary of the motif is shown in Figure 2A; the sites themselves are listed in Supplementary Table 1. To control for the possibility that including the 13 *H.influenzae* promoters had seeded the motif searches, we repeated the analysis with these promoters excluded; this identified the same motif at the same 74 sites in the other genomes.

The motif in Figure 2A resembles the CRP-S consensus, but more rigorous analysis required comparison with a dataset based on canonical CRP promoters. Thus we next determined whether the CRP-N sites in Sxy-independent *H.influenzae* promoters are also conserved in the other species. CRP-N sites regulate 41 transcriptional units in *H.influenzae*, encoding genes for sugar utilization, nutrient uptake and central metabolism during competence development (13). To provide comparable numbers of genes in the CRP-N and CRP-S datasets, we limited the CRP-N analysis to homologs from only *P.multocida, M.succiniciproducens* and *H.ducreyi*. This yielded one motif shared by 21 *M.succiniciproducens* sites, 35 *P.multocida* sites and 15 *H.ducreyi* sites (summarized by the sequence logo in Figure 2B; sites listed in Supplementary Table 2). As expected, this motif strongly resembled CRP-N sites.

The weight matrix method of Stormo and Hartzell (37) was used to quantify the similarities and differences between these putative CRP-S and CRP-N sites. We first scored all sites for goodness-of-fit with the 58 experimentally determined *H.influenzae* CRP-binding sites (CRP-N and CRP-S

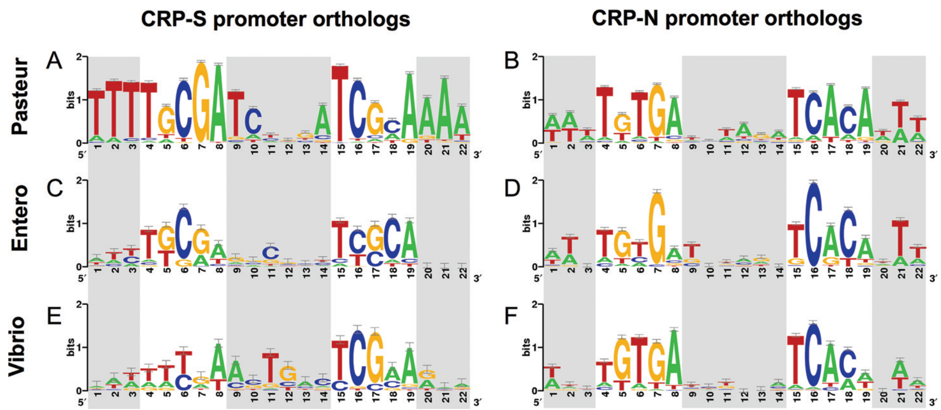

Figure 2. Motifs from pooled gene promoters. (**A** and **B**) *Pasteurellaceae*; (**C** and **D**) *Enterobacteriaceae*; (**E** and **F**) *Vibrionaceae*. CRP-S promoter orthologs are those in Figure 1. Logos were generated from alignment of all sites in Supplementary Tables 1 through 6 using WebLogo (http://weblogo.cbr.nrc.ca/logo.cgi). White bars highlight the conserved CRP-binding site motifs between positions 4–8 and 15–19. WebLogo employs a correction factor to compensate for underestimates of entropy arising from limited sequence data: error bars are twice the height of this correction (78).

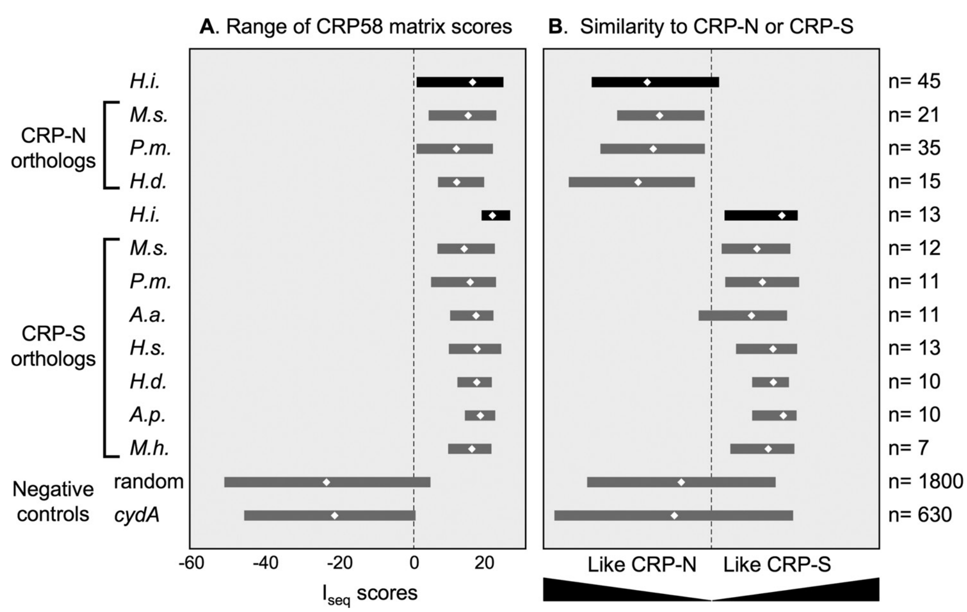

Figure 3. Similarity of putative CRP sites to experimentally determined sites (Black bars). Bars indicate range of scores; white diamonds are the mean scores. (**A**) Sites scored with CRP58 matrix. (**B**) Scores indicate the difference of I_{seq} for each site scored with CRP45 (CRP-N) and CRE13 (CRP-S) matrices.

combined). The weight scores (I_{seq}) for all sites overlapped the scores of the *H.influenzae* CRP sites used to construct the matrix (Figure 3A). The lowest two bars are controls, showing that all the predicted sites differ significantly from 1800 randomly generated sequences with the same G + C content as the average *Pasteurellaceaen* genome (40.4% G + C) and from all 22 bp sequences in the CRP-independent *cydA* promoter regions of *H.influenzae, M.succiniciproducens* and *P.multocida*. Sample means were compared using the Tukey-Kramer 'honestly significant difference' test for multiple-comparison of samples with unequal *n*. This confirmed that putative CRP-S and CRP-N sites are indistinguishable from one another when scored with the CRP58 matrix, but differ significantly from random and *cydA* sequence ($P < 0.0001$). These results indicate that all of the predicted CRP sites are very likely true CRP-binding sites.

To test whether the distinction between CRP-N and CRP-S sites exists in *Pasteurellaceae* other than *H.influenzae*, two more weight matrices were generated from subsets of the verified 58 *H.influenzae* CRP sites: one from the 13 CRP-S sites and the other from the 45 CRP-N sites. Figure 3B summarizes the scores of the *Pasteurellacean* promoters. All but one of the 74 predicted sites from *Pasteurellaceae* genes in Figure 1 (orthologs of *H.influenzae* CRP-S genes) scored higher with the CRP-S weight matrix than any of the CRP-N orthologs, with the sole exception of the *A.actinomycetemcomitans rec2* promoter site. Conversely, the 71 sites in all *M.succiniciproducens, P.multocida* and *H.ducreyi* orthologs of CRP-N-regulated genes scored higher with the CRP-N matrix. For all species, the CRP-S and CRP-N I_{seq} scores differ significantly (Tukey-Kramer, $P < 0.0001$). These results show that the CRP regulons are subdivided by CRP-S and CRP-N sites in all sequenced *Pasteurellaceae* genomes.

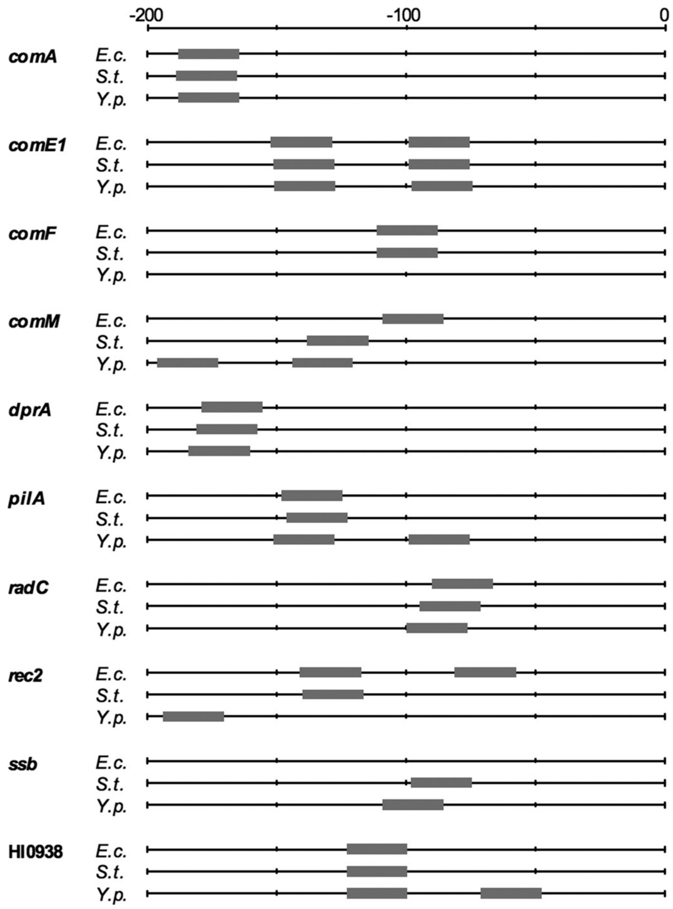

Figure 4. Physical map of *Enterobacteriaceae* promoters, named according to *H.influenzae* orthologs in Figure 1. Grey boxes indicate positions of putative CRP-S sites relative to start codons (sites listed in Supplementary Table 3). In all three *comA* promoters, a second CRP-S lies >200 bp away from the gene start (*E.c.* −246, *S.t.* −247, *Y.p.* −246).

CRP-S and CRP-N sites in the *Enterobacteriaceae*

Phylogenetic footprint analysis of the 33 *Enterobacteriaceae* promoters in Figure 1 (CRP-S orthologs) identified a single conserved motif present at 38 sites (summarized by the sequence logo in Figure 2C; sites are listed in Supplementary Table 3). Analyzing the 90 promoters of orthologs of *H.influenzae* CRP-N-regulated genes yielded 57 sites sharing one motif (sequence logo in Figure 2D; sites listed in Supplementary Table 4). As expected, the CRP-N-ortholog motif in Figure 2D is a canonical CRP site, whereas the CRP-S-ortholog promoter motif in Figure 2C has significant overrepresentation of the C_6 and G_{17} bases characteristic of CRP-S sites. Figure 4 shows physical maps of these predicted CRP-S promoters; for each gene the locations of putative CRP sites are often very similar in the three *Enterobacteriaceae*, providing further evidence of a conserved biological function. Taken together, these results are a strong indication that *Enterobacteriaceae* competence gene orthologs are part of a distinct regulon characterized by CRP-S sites. The lack of any previously characterized *Enterobacteriaceae* CRP-S sites precluded us from applying the weight-matrix analysis used for the *Pasteurellaceae* sites.

CRP-S and CRP-N sites in the *Vibrionaceae*

Although, *V.cholerae* had not been known to be naturally transformable, Meibom *et al.* (38) found that one of the two *V.cholerae sxy* orthologs, VC1153, and orthologs of *H.influenzae* competence genes *comA-E, pilA-D, pilF2* and *dprA* are among the genes induced when cells are cultured in the presence of chitin. They subsequently demonstrated that competence can be induced if cells are cultured with chitin (17), and that *sxy* is essential for competence, as in *H.influenzae*. Over-expression of the *sxy* ortholog VC1153 was also shown to up-regulate 99 genes, including the competence genes induced by chitin (17,38).

Consequently we expected to find CRP-S motifs in the promoters of the *H.influenzae* competence gene orthologs. However, when the 33 promoters from the *Vibrionaceae* species in Figure 1 were analysed as described for the *Enterobacteriaceae* and *Pasteurellaceae*, no significant conserved motifs were detected. Analyzing each species' promoters separately also failed to recover any significant motifs.

To narrow the set of genes being searched we used the *V.cholerae* gene expression studies. Analysis of the 78 promoters of the 99 Sxy-induced *V.cholerae* genes did not identify any significant shared motifs. However, the 99 Sxy-induced genes include six transcription factors and expression was not assayed until several cell-generations after induction of *sxy*, leading us to suspect that some of the 99 genes are not directly Sxy-regulated but induced secondarily by these other transcription factors. As some of the induced genes showed only modest induction and our analysis required high-confidence members of the Sxy regulon, we then limited our analysis to promoters induced by both Sxy and chitin (19 of 22 chitin-induced promoters, excluding *sxy* itself).

The three motif recognition algorithms agreed on a single motif shared by five promoters, *comA-F, pilA-D*, VC0047-*dprA, pilF2* and VCA0140. These five promoters were pooled with the homologous promoters from *V.parahaemolyticus* and *V.vulnificus*, and used for the motif search whose results are shown in Figure 2E. This search identified a single motif present at 24 sites in the 15 promoters (sites listed in Supplementary Table 5). The right half of the motif aligns well with the CRP-S motifs already found in *Enterobacteriaceae* and *Pasteurellaceae* promoters. Because the left half-motif only weakly resembles the CRP-S half-motif, the 19 *V.cholerae* promoters were re-examined for shorter motifs. This identified the motif 5′-ACTCG(A/C)AA in most of the 19 Sxy-induced *V.cholerae* promoters, but these shorter sites were excluded from further analysis because they were not consistently identified by all three search algorithms. However, all three algorithms scored this motif as more statistically significant than similar-sized motifs found in the other bacterial families. Because this short motif is contained within the sites summarized in Figure 2E, it appears to represent a shorter, more frequent variant of that longer motif.

The CRP-dependence of these genes has not been directly investigated, but natural transformation is catabolite repressed in *V.cholerae* (17), as expected for a CRP-dependent process. Taken together, these results strongly

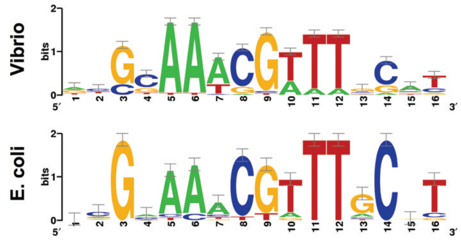

Figure 5. PurR logos from alignment of 27 *Vibrionaceae* sites in Supplementary Table 7 and the 15 *E.coli* sites listed at RegulonDB.

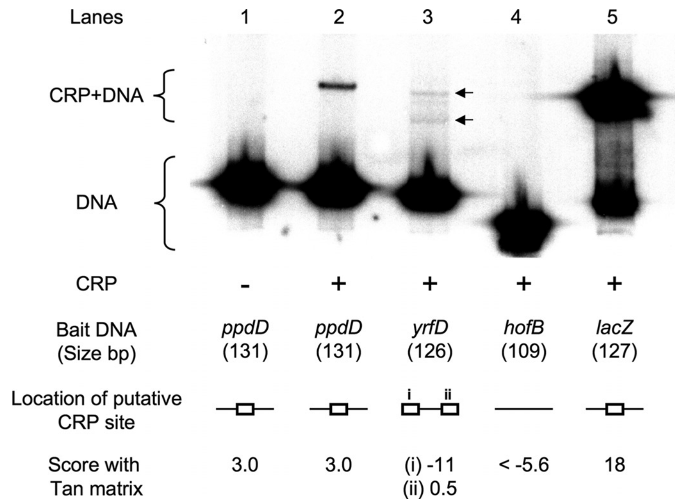

Figure 6. EMSA demonstrating site-specific binding by *E.coli* CRP to *E.coli* promoters containing putative CRP-S sites. Arrows indicate faint bandshifts with *yrfD* (b3395) promoter.

suggest that CRP-S sites mediate induction of natural competence in *V.cholerae* by CRP and Sxy.

Little is know about the global regulatory role of CRP in *Vibrionaceae*, where research has focused on the regulation of virulence (39,40). To determine whether CRP regulates a similar set of genes to those seen in the *Enterobacteriaceae* and *Pasteurellaceae*, we examined promoters of orthologs of *H.influenzae* CRP-N-regulated genes for shared motifs. This analysis found two highly conserved motifs: the expected one matching the CRP sites found in the *Enterobacteriaceae* and *Pasteurellaceae* (Figure 2F), and one matching the PurR repressor binding site consensus (Figure 5); the genes and sites are listed in Supplementary Tables 6 and 7. The CRP motif in Figure 2F shows very strong overrepresentation of T:A_6 and A:T_{17}, placing these sites in the CRP-N regulon as in *Pasteurellaceae* and *Enterobacteriaceae*.

PurR represses nucleotide biosynthesis genes when intracellular purine nucleotide pools are high. The candidate PurR sites were detected in 24 of the 71 *Vibrio* promoters (8 in *V.cholerae*, 7 in *V.parahaemolyticus*, and 9 in *V.vulnificus*), including 13 of those that also had CRP-N motifs (Supplementary Table 7). Of the eight *V.cholerae* promoters, two (*purE* and *uraA*) are members of the PurR regulon predicted by TractorDB and by Ravcheev *et al.* (41), and are also regulated by both CRP and PurR in *H.influenzae* (13). This analysis adds six new promoters (*cdd, fbp, mdh, mglB, rbsD* and *pckA*) to the 19 previously predicted promoters in the *V.cholerae* PurR regulon. Two of these six (*cdd* and *rbsD*) regulate genes involved in nucleotide metabolism, so their inclusion in the PurR regulon is not surprising. The remaining four promoters regulate galactose uptake genes (*mglB*) and genes for synthesizing precursor metabolites (*fbp, mdh* and *pckA*).

Pseudomonadaceae and *Xanthomonadaceae* orthologs lack conserved regulatory motifs

Although none of the *Pseudomonadaceae* and *Xanthomonadaceae* genomes listed in Figure 1 contained *sxy* orthologs, CRP orthologs are present. In *Pseudomonadaceae*, the CRP ortholog Vfr (virulence factor regulator) regulates quorum sensing, protein secretion, motility and adherence (42–45). In *Xanthomonadaceae*, the CRP ortholog Clp (CAP-like protein) regulates the synthesis of extracellular enzymes, pigment and xanthum gum (46,47).

Because significantly fewer *H.influenzae* CRP-N genes are conserved in the *Pseudomonadaceae* and *Xanthomonadaceae* than in other families, we searched five additional genomes of each family for homologs of *H.influenzae* genes with CRP-N and CRP-S sites (Table 1). For each family the genomes used are specified in Materials and Methods.

We identified 63 *Pseudomonadaceae*-promoter and 68 *Xanthomonadaceae*-promoter orthologs of *H.influenzae* CRP-S-regulated genes. No conserved motifs were detected in the promoters from either family. Transcriptome analysis has found that Vfr weakly induces members of the *pilM-Q* (*comA-E* orthologs) and *pilB-D* operons, in addition to many genes involved in motility and adherence (45). However, motif searches restricted to the *pilM-Q* and *pilB-D* promoters in all *Pseudomonadaceae* did not identify any conserved motif. In the absence of expression data for Clp in *Xanthomonadaceae*, we could not further refine our search parameters.

We similarly analysed 119 *Pseudomonadaceae*-promoter and 77 *Xanthomonadaceae*-promoter orthologs of *H.influenzae* CRP-N-regulated genes. Neither pool of promoters contained a significant conserved motif. The absence of conserved motifs suggests that orthologs of *H.influenzae* CRP-regulated genes are not CRP-regulated in the *Pseudomonadaceae* or *Xanthomonadaceae*.

Regulation of predicted *E.coli* CRP-S promoters by CRP and Sxy

The above bioinformatics analysis suggested that the extensive experimental work on CRP function in *E.coli* has overlooked the Sxy-specific CRP sites. We directly tested the regulation of these sites in *E.coli*.

First, to test whether CRP binds specifically to *E.coli* CRP-S sites, we purified His-tagged *E.coli* CRP under native conditions and used electrophoretic mobility-shift assays (EMSA) to detect site-specific DNA-binding (Figure 6). We tested binding to the *E.coli ppdD* (b0108; *pilA* ortholog) and *yrfD* (b3395; *comA* ortholog) promoters, which contain one and two predicted CRP-S sites respectively but no predicted CRP-N sites. The *E.coli lacZ* promoter served as a positive control as it contains a well-studied CRP-binding site. The *hofB* (b0109; *pilB* homolog) gene is adjacent to *ppdD* but does not contain any CRP site; it and cloning-vector DNA (data not shown) served as negative controls. No bandshifts were observed in the absence of CRP or with negative control *hofB* DNA (Figure 6, lanes 1 and 4). Bandshifts are apparent in lanes 2 and 3, although very little DNA is shifted relative to the *lacZ* promoter in lane 5, indicating that CRP has low but specific affinity for CRP-S sites. The *yrfD* promoter generates two faint bands; the higher molecular weight band is likely the result of occupancy of both CRP sites, the lower molecular band from CRP binding to only one site. The greater mobility of these *yrfD*-CRP promoter complexes relative to *ppdD* and *lacZ* complexes may be because the CRP-S sites are at the ends of the *yrfD* DNA fragment (indicated in Figure 6)—CRP-induced DNA bending is known to reduce mobility in these assays, and the effect is smaller if the CRP site is near the fragment's end (11,48). For each site the I_{seq} scores generated from the standard *E.coli* CRP weight matrix (4) are shown at the bottom of Figure 6; the low affinity of CRP for the *ppdD* and *yrfD* CRP-S sites is consistent with their low scores.

Having found that the predicted CRP-S sites in *E.coli* are bona fide, albeit weak, CRP sites, we used quantitative PCR to test whether two of the *E.coli* genes with CRP-S promoters (*ppdD* and *yrfD*) are CRP-induced *in vivo*, and whether this induction is Sxy-dependent. The *E.coli sbmC* gene was included in this analysis; it has no *H.influenzae* homolog but is CRP regulated, and its predicted CRP site resembles the CRP-S motif (3) (Figure 7A). A representative CRP-N-regulated gene, *mglB,* was also included. To examine Sxy dependence, exponentially growing cells carrying *E.coli sxy* cloned under LacI repression were induced with IPTG (Figure 7B and C). The red bars in Figure 7B show that IPTG induction of Sxy induced *ppdD* 90-fold, *yrfD* 16-fold, and *sbmC* 6-fold, but had no detectable effect on *mglB*. Previous studies have found that the *E.coli yrfD-hofQ* operon is transcribed either poorly or undetectably [summarized by (49)]; attempts to detect *ppdD* transcript have also failed (50). This is the first demonstration that these genes are not only transcribed but very strongly induced by Sxy. These findings also imply that the amount of Sxy in LB-grown cells is too low to permit induction of *yrfD* and *ppdD*.

To test the CRP-dependence of these genes, transcription analysis was repeated using a host carrying a *crp* knockout (Figure 7B). Comparison of the grey and green bars shows that induction of all four genes is absolutely dependent on CRP, confirming the bandshift results. Because Sxy is thought not to bind DNA, we also examined gene expression in $cyaA^-$ cells to test whether Sxy might act by overriding CRP's dependence on its allosteric effector cAMP. In this genetic background, exogenous cAMP was required for induction of all four genes (Figure 7C), indicating that Sxy does not bypass CRP's cAMP-dependence. Again, whereas induction of *ppdD, yrfD* and *sbmC* absolutely required both CRP and Sxy, *mglB* was induced by exogenous cAMP to the same levels in the presence or absence of Sxy. All four genes were also catabolite repressed by the addition of glucose to culture medium, and induction was restored upon addition of cAMP (data not shown). Together, these results indicate that *E.coli* CRP-S promoters are genuine CRP-dependent promoters, and that they are Sxy-dependent, as in *H.influenzae*. Because *sbmC's* Sxy dependence was

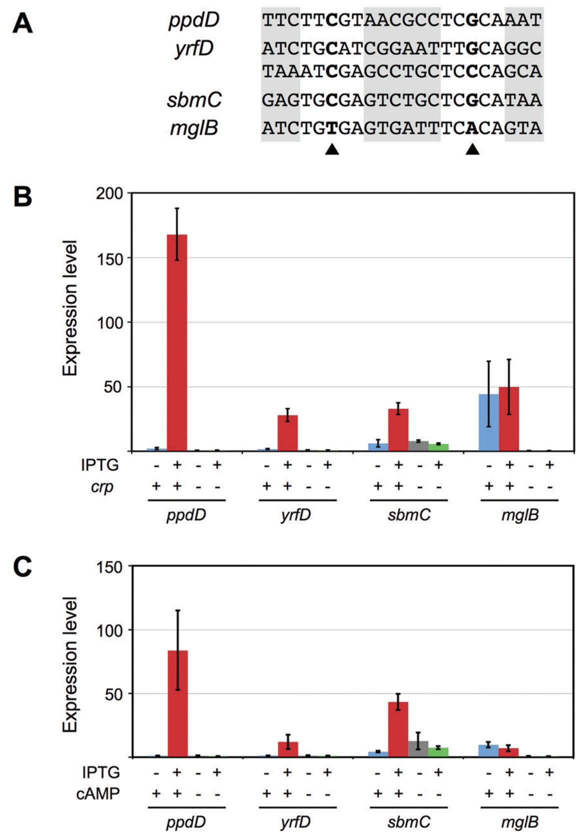

Figure 7. Sxy-dependent gene expression from *E.coli* CRP-S promoters measured using quantitative PCR. (**A**) Alignment of CRP sites; arrows highlight positions 6 and 17. (**B**) Gene expression in wild-type and crp^- cells carrying cloned, IPTG-inducible *E.coli sxy* (pASKA*sxy*). (**C**) Gene expression in $cyaA^-$ cells carrying pASKA*sxy*. The average and standard deviation of two or more independent cultures are shown, and expression levels are expressed as 1/1000 of 23S rRNA abundance.

predicted only by its CRP-S motif, they also validate use of this motif as a predictor of Sxy dependence.

E.coli cells carrying a plasmid expressing *H.influenzae sxy* had substantially elevated levels of *ppdD, yrfD* and *sbmC* but not *mglB* compared to cells with a control plasmid (data not shown). This implies that Sxy's as-yet-uncharacterized mode of action is the same in *E.coli* and *H.influenzae*, and is consistent with previous work showing that *E.coli* CRP fully complements a *H.influenzae crp* mutant for competence induction (51).

Evolution of CRP and Sxy in γ-proteobacteria

The above analysis revealed that specialized CRP sites regulate competence genes in the *Enterobacteriaceae, Pasteurellaceae* and *Vibrionaceae* (the 'EPV' clade), but not in the *Pseudomonadaceae* or *Xanthomonadaceae*. We used phylogenetic analysis to look for specific features of CRP that evolved in the EPV clade to allow interaction with Sxy or CRP-S sites. In examining CRP-FNR protein evolution, Korner *et al.* (52) have shown that γ-proteobacterial CRP proteins constitute a monophyletic clade, distantly related to other CRP-FNR proteins in eubacteria. However, this analysis had little resolution within the

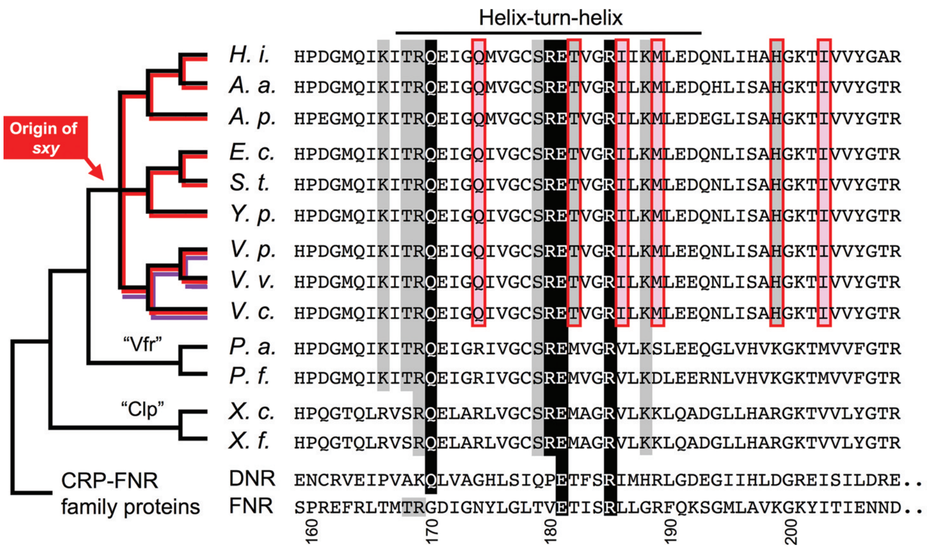

Figure 8. Evolutionary history of *crp* (black) and *sxy* (red). Nodes of the phylogenetic tree are supported by bootstrap values over 80%, except for the root of the EPV clade where branching order could not be resolved due to low (<40%) bootstrap values. A gene duplication event generated *sxy* paralogs in the *Vibrionaceae* (red and purple branches). CRP DNA-binding domains are aligned with the closely related *P.aeruginosa* DNR and the distantly related *E.coli* FNR. Amino acids in *E.coli* CRP that make base contacts are highlighted black, those making contact with phosphates in the DNA backbone are highlighted grey, those shared only by the EPV clade are outlined in red; amino acid numbering is according to *E.coli* CRP. Species names as in Figure 1.

EPV clade and its results disagreed with the established relationships presented in Figure 1. We reconstructed CRP evolution with a narrower focus, restricting the analysis to the CRP orthologs of the five families we have examined (shown in Figure 8). Five lineages were resolved, and their congruence with the established bacterial phylogeny shown on the left of Figure 1 confirms the findings of Korner *et al*. (52) that CRP is ancestral to the γ-proteobacteria. The Sxy phylogeny in Figure 8 is also congruent with the established phylogeny for the EPV clade, supporting the null hypothesis that neither Sxy nor CRP has been transferred horizontally between species.

Three amino acids in the *E.coli* CRP helix–turn–helix confer CRP-N site recognition through base contacts (R180, E181 and R185); Figure 8 shows that they are conserved in all five families. Q170 is also conserved; it makes a base contact, but its contribution to DNA site specificity has not been investigated (53). Consistent with conservation of these amino acids, CRPs from *E.coli, H.influenzae* and *X.campestris* preferentially bind the motif $T_4G_5T_6G_7A_8$ (54,55) (A. Cameron, manuscript in preparation)—comparable binding experiments have not yet been done for CRP in other families. Thus, specificity for the CRP-N motif evolved before the last γ-proteobacterial common ancestor.

CRP's DNA-binding domain is contained within 50 C-terminal amino acids (aligned in Figure 8); six of these are shared only within the EPV clade, as expected of residues that might mediate interactions with CRP-S sites. Nothing is known about Q174, I186, M189 or I203, but both T182 and H199 contribute to DNA binding in *E.coli*. H199 is particularly intriguing because, along with K26 and K166, it induces a secondary, stabilizing kink in CRP-binding sites through contacts with phosphates at positions 1–3 and 20–22 (53). The absence of H199 from *Pseudomonadaceae* and of all three residues from *Xanthomonadaceae* suggests that the secondary kink may be less important in these two families. Because CRP-S sequences hinder primary kink formation, the secondary kink may play a key role at CRP-S sites, especially in the *Pasteurellaceae* where CRP-S sites have a dramatic overrepresentation of flexible A and T runs at positions 1–3 and 20–22 (Figure 2A). Thus, we postulate that both the CRP-S motif and Sxy arose in the EPV common ancestor, and that this coincided with the introduction of H199 to strengthen the secondary DNA kink.

DISCUSSION

We have identified in many of the γ-proteobacteria a mode of CRP regulation that initially was known only for the competence genes of *H.influenzae*. Most notably, in *E.coli* CRP binds to and stimulates transcription at novel CRP sites with a distinct consensus (CRP-S) that makes transcription activation dependent on an additional protein factor, Sxy. The analysis also extended evidence for natural competence to the five best-known γ-proteobacterial families.

The mechanism by which Sxy facilitates CRP–DNA interactions is not known. However a wealth of information is available about how other factors affect CRP-regulated promoters in *E.coli*; Figure 9 summarizes these. Promoters such as *lacZYA*, where CRP is the sole activator, have high-affinity CRP sites; here CRP makes protein contacts only with RNA polymerase. At slightly more complex promoters, such as *proP* and *malE*, CRP and other transcription factors bind independently to high-affinity sites in promoter DNA, but act synergistically to recruit RNA polymerase (56,57). At promoters where CRP binds cooperatively with other proteins, higher-order nucleoprotein complexes form. For example, CRP depends on direct protein–protein interactions with MelR and CytR to bind low-affinity CRP sites in the *melA* and *deoC* promoters, respectively (58–60).

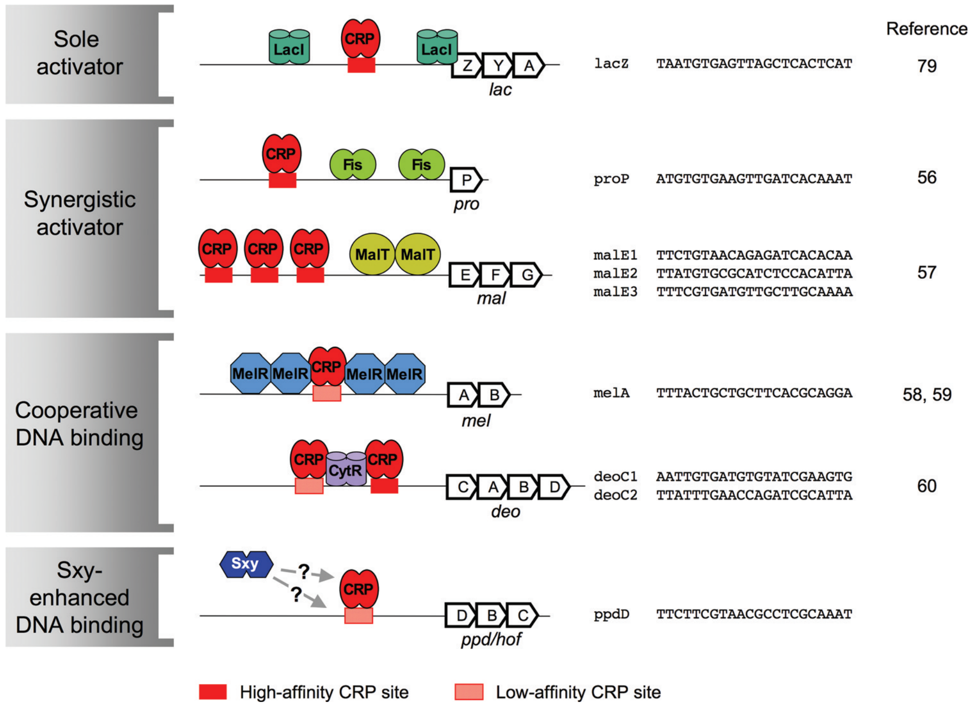

Figure 9. Categories of CRP-activated promoters in *E.coli*. In promoters where CRP acts as a sole transcription activator or synergistic activator, it binds to high-affinity sites. CRP requires cooperativity with MelR and CytR to bind to low affinity sites in the *melAB* and *deoCABD* promoters, among others. Sxy is hypothesized to interact directly with CRP, and may also bind DNA.

CRP-S promoters are distinctive in having no apparent shared binding sites for Sxy or other factors. This is consistent with the absence of recognizable DNA binding motifs in Sxy itself. We hypothesize that Sxy interacts with CRP to stabilize CRP-DNA binding, possibly by reducing the free energy requirements for DNA kinking between the unfavourable C_6–G_7 base pairs. This predicts that Sxy should enhance bandshifting by CRP at CRP-S sites. Unfortunately, our ongoing experiments to test this have been hindered by Sxy's poor stability in expression cultures.

There are two classes of CRP-dependent promoters [reviewed in (61) and in (8)]. At class I promoters, such as *lac*, CRP binding sites are located near −62, −72, −83 or −93 relative to the transcription start site. When CRP binds to these sites, its activating region 1 (AR1) contacts RNAP's α subunit C-terminal domain (αCTD) to recruit RNAP to the promoter. At class II promoters, CRP binds near −42 and contacts occur between CRP's AR1, AR2 and AR3 and the RNAP αCTD, αNTD and σ subunits, respectively. *H.influenzae* CRP-S sites are located near −62, −73 and −100 (13), placing them in class I, while the *E.coli sbmC* (CRP-S) promoter has been shown to operate through a class I mechanism (3). We expect all CRP-S promoters to belong to class I. Consequently, CRP will not be intimately associated with RNAP at these promoters, leaving more regions of the protein exposed for possible interactions with Sxy.

Might Sxy also enhance CRP activation at sites that do not fit the CRP-S consensus? In many of the CRP sites that regulate orthologs of Sxy-dependent *H.influenzae* genes, only one half site has the C_6 base of the CRP-S consensus, as does the Sxy-dependent *H.influenzae* HI1631 promoter. In the *Pasteurellaceae*, the second half-site rarely matches the CRP-N consensus, and we predict that these sites will also be Sxy-dependent. In all *Enterobacteriaceae* CRP-S ortholog promoters, half sites that do not have C_6 always have G_6 (Supplementary Table 3). Thus, *Enterobacteriaceae* CRP-S sites are striking in never having the T_6 characteristic of the CRP-N motif. We do not know whether Sxy will also enhance CRP activation of the 39 (out of 182) *E.coli* CRP sites in RegulonDB that have the CRP-S C_6 base in one-half site but not the other (e.g. the *melA* and *deoC* sites in Figure 9).

Although, the competence genes in the CRP-S regulon are ancestral to the EPV clade, the CRP-S sites that regulate them are likely to be dynamic, decaying and arising anew. For example, two CRP-S sites are predicted in each of the *Enterobacteriaceae comA-E* and *comE1* promoters, unlike the single sites in each *Pasteurellaceae* promoter (Figure 4). *comF* has its own promoter in *Enterobacteriaceae, Vibrionaceae* and most *Pasteurellaceae* species, but not in *H.influenzae* where it has joined the *comA-E* operon to retain CRP-S regulation (Figure 1). Moreover, in *H.somnus* the *pil* operon has dissociated into three transcription units: *pilA, pilB, pilCD*, each with its own CRP-S site. This indicates that these genes are under strong selection to maintain CRP-S regulation.

Almost all of the genes in the *H.influenzae* competence regulon are conserved throughout the γ-proteobacteria (Figure 1). Most of these are known to function in DNA binding and transport across the outer and inner membranes, but others encode cytoplasmic proteins (SSB, RadC, SbmC, DprA and ComM). Although some of the latter may be induced to promote recombination, consideration of the evolutionary function of competence may help explain both the signaling role of Sxy and the inclusion of cytoplasmic proteins in its regulon.

The most immediate consequence of DNA uptake is the provision of nucleotides, both from the strand brought into the cytoplasm and from the strand degraded at the cell surface (Gram positives) or in the periplasm (Gram negatives) (62). Nucleotide depletion is known to be necessary for competence induction in *H.influenzae* (63), and our preliminary experiments indicate that this is mediated by induction of Sxy (A. Cameron, manuscript in preparation). Thus Sxy may serve as a signal of nucleotide depletion.

This role for Sxy suggests that the CRP-S regulons may have functions beyond that of DNA uptake. In particular, depletion of intracellular nucleotide pools threatens chromosome integrity by causing replication forks to stall. *E.coli* employs several strategies to reduce the deleterious effects of stalled replication [reviewed in (64)], and the cytoplasmic CRP-S-regulated genes have cellular roles that can contribute to these. SSB binds to ssDNA at stalled and aborted replication forks to reinitiate replication by helping reload the replisome (65). RadC facilitates recombinational repair at stalled replication forks (66). SbmC (also called GyrI) specifically inhibits DNA gyrase and consequently blocks gyrase-mediated DNA lesions during replication (67,68). DprA protects ssDNA from degradation in *Streptococcus pneumoniae* (69), and imported DNA is rapidly degraded in *H.influenzae* cells lacking DprA or ComM (70,71). *comM* (b3765) is induced by ultraviolet (UV) irradiation (72), further supporting a role in maintaining chromosome integrity. To summarize, the CRP-S regulon may unite genes that alleviate problems arising from depleted nucleotide pools; competence proteins scavenge extracellular DNA while cytoplasmic proteins protect ssDNA and promote recombination in order to resolve stalled replication forks.

More generally, the 'nutritional competence' demonstrated in *E.coli* is likely the best model for the role of DNA uptake in bacteria (21,49). Palchevskiy and Finkel (49) have shown that *com* genes enable *E.coli* to grow with DNA as the sole nutrient source and that this ability is important in long-term culture. Other bacteria may also benefit from using DNA as a nutrient, as it is abundant in many natural environments. DNA concentrations of several 100 μg/ml are typical in the mammalian mucosal niches utilized by *E.coli* and *H.influenzae* (73). In fact, DNA's stability after cell death and lysis causes it to accumulate in many of the aquatic, soil, and animal/plant host niches inhabited by γ-proteobacteria (74). This extracellular DNA is nutritionally significant; in marine sediments it provides prokaryotes with 4% of their carbon, 7% of their nitrogen and nearly 50% of their phosphate (75).

The 13 CRP-S sites we found in *E.coli* promoters have been overlooked in earlier genome-wide searches because they score very low with weight matrices derived from canonical *E.coli* CRP (CRP-N) sites (Figure 6) (4,76,77). We detected these unusual sites using orthology information to identify candidate promoters, and then accepted only sites selected by all of three motif recognition algorithms. The stringency of our bioinformatics approach means that it almost certainly will have missed some CRP-S sites. The true extent of the CRP-S regulons in different bacteria will be readily revealed by global transcriptome analysis using both Sxy and CRP mutants, like that done in *H.influenzae*. The true extent of competence in the γ-proteobacteria may be harder to determine, as conditions that induce these regulons are not yet understood.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Jacques van Helden for developing the excellent, user-friendly RSATools, and Brett McLeod and Steve Seredick for help with bandshift assays. The authors are grateful to an anonymous reviewer for helpful suggestions on CRP-DNA interactions, and the authors also wish to thank members of the Redfield lab for useful discussion and feedback. *E.coli* strains were kindly provided by GenoBase and by Tyrrell Conway. The authors thank TIGR for providing a valuable genome analysis resource. Funding for this work was provided to R.J.R. through the Canadian Institute of Health Research. Preliminary *M.haemolytica* sequence data were obtained from the Baylor College of Medicine Human Genome Sequencing Center website (http://www.hgsc.bcm.tmc.edu); USDA/NRICGP grant 00- 35204-9229 to Sarah Highlander and George Weinstock at the BCM-HGSC support this project. *A.actinomycetemcomitans* sequence data were provided by the Actinobacillus Genome Sequencing Project (Bruce A. Roe, Fares Z. Najar, Allison Gillaspy, Sandra Clifton, Tom Ducey, Lisa Lewis and D.W. Dyer); this project is supported by a USPHS/NIH grant from the National Institute of Dental Research. Funding to pay the Open Access publication charges for this article was provided by a grant to R.J.R. from the Canadian Institutes for Health Research.

Conflict of interest statement. None declared.

REFERENCES

1. Emmer,M., de Crombrugghe,B., Pastan,I. and Perlman,R. (1970) Cyclic AMP receptor protein of *E. coli*: its role in the synthesis of inducible enzymes. *Proc. Natl Acad. Sci. USA*, **66**, 480–487.
2. McKay,D.B. and Steitz,T.A. (1981) Structure of catabolite gene activator protein at 2.9 A resolution suggests binding to left-handed B-DNA. *Nature*, **290**, 744–749.
3. Zheng,D., Constantinidou,C., Hobman,J.L. and Minchin,S.D. (2004) Identification of the CRP regulon using *in vitro* and *in vivo* transcriptional profiling. *Nucleic Acids Res.*, **32**, 5874–5893.
4. Tan,K., Moreno-Hagelsieb,G., Collado-Vides,J. and Stormo,G.D. (2001) A comparative genomics approach to prediction of new members of regulons. *Genome Res.*, **11**, 566–584.
5. Schultz,S.C., Shields,G.C. and Steitz,T.A. (1991) Crystal structure of a CAP-DNA complex: the DNA is bent by 90 degrees. *Science*, **253**, 1001–1007.
6. Chen,S., Gunasekera,A., Zhang,X., Kunkel,T.A., Ebright,R.H. and Berman,H.M. (2001) Indirect readout of DNA sequence at the primary-kink site in the CAP- DNA complex: alteration of DNA binding specificity through alteration of DNA kinking. *J. Mol. Biol.*, **314**, 75–82.
7. Chen,S., Vojtechovsky,J., Parkinson,G.N., Ebright,R.H. and Berman,H.M. (2001) Indirect readout of DNA sequence at the

primary-kink site in the CAP–DNA complex: DNA binding specificity based on energetics of DNA kinking. *J. Mol. Biol.*, **314**, 63–74.
8. Lawson,C.L., Swigon,D., Murakami,K.S., Darst,S.A., Berman,H.M. and Ebright,R.H. (2004) Catabolite activator protein: DNA binding and transcription activation. *Curr. Opin. Struct. Biol.*, **14**, 10–20.
9. Barnard,A., Wolfe,A. and 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.
10. Gaston,K., Kolb,A. and Busby,S. (1989) Binding of the *Escherichia coli* cyclic AMP receptor protein to DNA fragments containing consensus nucleotide sequences. *Biochem. J.*, **261**, 649–653.
11. Kolb,A., Spassky,A., Chapon,C., Blazy,B. and Buc,H. (1983) On the different binding affinities of CRP at the lac, gal and malT promoter regions. *Nucleic Acids Res.*, **11**, 7833–7852.
12. Pyles,E.A. and Lee,J.C. (1996) Mode of selectivity in cyclic AMP receptor protein-dependent promoters in *Escherichia coli*. *Biochemistry*, **35**, 1162–1172.
13. Redfield,R.J., Cameron,A.D., Qian,Q., Hinds,J., Ali,T.R., Kroll,J.S. and Langford,P.R. (2005) A novel CRP-dependent regulon controls expression of competence genes in *Haemophilus influenzae*. *J. Mol. Biol.*, **347**, 735–747.
14. Macfadyen,L.P. (2000) Regulation of competence development in *Haemophilus influenzae*. *J. Theor. Biol.*, **207**, 349–359.
15. Williams,P.M., Bannister,L.A. and Redfield,R.J. (1994) The *Haemophilus influenzae* sxy-1 mutation is in a newly identified gene essential for competence. *J. Bacteriol.*, **176**, 6789–6794.
16. Redfield,R.J., Findlay,W.A., Bosse,J., Kroll,J.S., Cameron,A.D.S. and Nash,J.H.E. (2006) Evolution of competence and DNA uptake specificity in the Pasteurellaceae. *BMC Evol. Biol.,* in press.
17. Meibom,K.L., Blokesch,M., Dolganov,N.A., Wu,C.Y. and Schoolnik,G.K. (2005) Chitin induces natural competence in Vibrio cholerae. *Science*, **310**, 1824–1827.
18. Carlson,C.A., Pierson,L.S., Rosen,J.J. and Ingraham,J.L. (1983) *Pseudomonas stutzeri* and related species undergo natural transformation. *J. Bacteriol.*, **153**, 93–99.
19. Tomb,J.F., el-Hajj,H. and Smith,H.O. (1991) Nucleotide sequence of a cluster of genes involved in the transformation of *Haemophilus influenzae* Rd. *Gene*, **104**, 1–10.
20. VanWagoner,T.M., Whitby,P.W., Morton,D.J., Seale,T.W. and Stull,T.L. (2004) Characterization of three new competence-regulated operons in *Haemophilus influenzae*. *J. Bacteriol.*, **186**, 6409–6421.
21. Finkel,S.E. and Kolter,R. (2001) DNA as a nutrient: novel role for bacterial competence gene homologs. *J. Bacteriol.*, **183**, 6288–6293.
22. van Helden,J. (2003) Regulatory sequence analysis tools. *Nucleic Acids Res.*, **31**, 3593–3596.
23. Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.
24. Neuwald,A.F., Liu,J.S. and Lawrence,C.E. (1995) Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci.*, **4**, 1618–1632.
25. Liu,X., Brutlag,D.L. and Liu,J.S. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.*, **6**, 127–138.
26. Hu,J., Li,B. and Kihara,D. (2005) Limitations and potentials of current motif discovery algorithms. *Nucleic Acids Res.*, **33**, 4899–4913.
27. Kitagawa,M., Ara,T., Arifuzzaman,M., Ioka-Nakamichi,T., Inamoto,E., Toyonaga,H. and Mori,H. (2005) Complete set of ORF clones of *Escherichia coli* ASKA library (A complete set of *E. coli* K-12 ORF archive). Unique resources for biological research. *DNA Res.*, **12**, 291–299.
28. Baba,T., Ara,T., Hasegawa,M., Takai,Y., Okumura,Y., Baba,M., Datsenko,K.A., Tomita,M., Wanner,B.L. and Mori,H. (2006) Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.*, **2**, E1–E11.
29. Seidman,E.C. (1995) Introduction of plasmid DNA into cells. In Ausubel,F.M. (ed.), *Current Protocols in Molecular Biology*. Brooklyn, NY, Vol. I, pp. 1.8.1–1.8.3.
30. Peekhaus,N. and Conway,T. (1998) Positive and negative transcriptional regulation of the *Escherichia coli* gluconate regulon gene gntT by GntR and the cyclic AMP (cAMP)-cAMP receptor protein complex. *J. Bacteriol.*, **180**, 1777–1785.
31. Hall,B.G. (2004) *Phylogenetic Trees Made Easy: A How-To Manual*. Sinauer Associates, Inc, Sunderland, Massachusetts, USA, pp. 154–156.
32. Falsenstein,J. (1989) PHYLIP—Phylogeny Inference Package (Version 3.2). *Cladistics*, **5**, 164–166.
33. Lerat,E., Daubin,V. and Moran,N.A. (2003) From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-Proteobacteria. *PLoS. Biol.*, **1**, 101–109.
34. Belda,E., Moya,A. and Silva,F.J. (2005) Genome rearrangement distances and gene order phylogeny in gamma-Proteobacteria. *Mol. Biol. Evol.*, **22**, 1456–1467.
35. Ciccarelli,F.D., Doerks,T., von Mering,C., Creevey,C.J., Snel,B. and Bork,P. (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science*, **311**, 1283–1287.
36. Price,M.N., Huang,K.H., Arkin,A.P. and Alm,E.J. (2005) Operon formation is driven by co-regulation and not by horizontal gene transfer. *Genome Res.*, **15**, 809–819.
37. Stormo,G.D. and Hartzell,G.W., 3rd (1989) Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl Acad. Sci. USA*, **86**, 1183–1187.
38. Meibom,K.L., Li,X.B., Nielsen,A.T., Wu,C.Y., Roseman,S. and Schoolnik,G.K. (2004) The *Vibrio cholerae* chitin utilization program. *Proc. Natl Acad. Sci. USA*, **101**, 2524–2529.
39. Choi,M.H., Sun,H.Y., Park,R.Y., Kim,C.M., Bai,Y.H., Kim,Y.R., Rhee,J.H. and Shin,S.H. (2006) Effect of the crp mutation on the utilization of transferrin-bound iron by *Vibrio vulnificus*. *FEMS Microbiol. Lett.*, **257**, 285–292.
40. Skorupski,K. and Taylor,R.K. (1997) Cyclic AMP and its receptor protein negatively regulate the coordinate expression of cholera toxin and toxin-coregulated pilus in *Vibrio cholerae*. *Proc. Natl Acad. Sci. USA*, **94**, 265–270.
41. Ravcheev,D.A., Gel'fand,M.S., Mironov,A.A. and Rakhmaninova,A.B. (2002) [Purine regulon of gamma-proteobacteria: a detailed description]. *Genetika*, **38**, 1203–1214.
42. Smith,R.S., Wolfgang,M.C. and Lory,S. (2004) An adenylate cyclase-controlled signaling network regulates *Pseudomonas aeruginosa* virulence in a mouse model of acute pneumonia. *Infect. Immun.*, **72**, 1677–1684.
43. Suh,S.J., Runyen-Janecky,L.J., Maleniak,T.C., Hager,P., MacGregor,C.H., Zielinski-Mozny,N.A., Phibbs,P.V., Jr and West,S.E. (2002) Effect of vfr mutation on global gene expression and catabolite repression control of *Pseudomonas aeruginosa*. *Microbiology*, **148**, 1561–1569.
44. Albus,A.M., Pesci,E.C., Runyen-Janecky,L.J., West,S.E. and Iglewski,B.H. (1997) Vfr controls quorum sensing in *Pseudomonas aeruginosa*. *J. Bacteriol.*, **179**, 3928–3935.
45. Wolfgang,M.C., Lee,V.T., Gilmore,M.E. and Lory,S. (2003) Coordinate regulation of bacterial virulence genes by a novel adenylate cyclase-dependent signaling pathway. *Dev. Cell*, **4**, 253–263.
46. de Crecy-Lagard,V., Glaser,P., Lejeune,P., Sismeiro,O., Barber,C.E., Daniels,M.J. and Danchin,A. (1990) A *Xanthomonas campestris pv. campestris* protein similar to catabolite activation factor is involved in regulation of phytopathogenicity. *J. Bacteriol.*, **172**, 5877–5883.
47. Kobayashi,D.Y., Reedy,R.M., Palumbo,J.D., Zhou,J.M. and Yuen,G.Y. (2005) A clp gene homologue belonging to the Crp gene family globally regulates lytic enzyme production, antimicrobial activity, and biological control activity expressed by Lysobacter enzymogenes strain C3. *Appl. Environ. Microbiol.*, **71**, 261–269.
48. Bai,G., McCue,L.A. and McDonough,K.A. (2005) Characterization of *Mycobacterium tuberculosis* Rv3676 (CRPMt), a cyclic AMP receptor protein-like DNA binding protein. *J. Bacteriol.*, **187**, 7795–7804.
49. Palchevskiy,V. and Finkel,S.E. (2006) *Escherichia coli* competence gene homologs are essential for competitive fitness and the use of DNA as a nutrient. *J. Bacteriol.*, **188**, 3902–3910.
50. Sauvonnet,N., Gounon,P. and Pugsley,A.P. (2000) PpdD type IV pilin of *Escherichia coli* K-12 can Be assembled into pili in *Pseudomonas aeruginosa*. *J. Bacteriol.*, **182**, 848–854.
51. Chandler,M.S. (1992) The gene encoding cAMP receptor protein is required for competence development in *Haemophilus influenzae* Rd. *Proc. Natl Acad. Sci. USA*, **89**, 1626–1630.
52. Korner,H., Sofia,H.J. and Zumft,W.G. (2003) Phylogeny of the bacterial superfamily of Crp-Fnr transcription regulators: exploiting the metabolic spectrum by controlling alternative gene programs. *FEMS Microbiol. Rev.*, **27**, 559–592.

53. Parkinson,G., Wilson,C., Gunasekera,A., Ebright,Y.W., Ebright,R.E. and Berman,H.M. (1996) Structure of the CAP-DNA complex at 2.5 angstroms resolution: a complete picture of the protein-DNA interface. *J. Mol. Biol.*, **260**, 395–408.
54. Gunasekera,A., Ebright,Y.W. and Ebright,R.H. (1992) DNA sequence determinants for binding of the *Escherichia coli* catabolite gene activator protein. *J. Biol. Chem.*, **267**, 14713–14720.
55. Dong,Q. and Ebright,R.H. (1992) DNA binding specificity and sequence of *Xanthomonas campestris* catabolite gene activator protein-like protein. *J. Bacteriol.*, **174**, 5457–5461.
56. McLeod,S.M., Aiyar,S.E., Gourse,R.L. and Johnson,R.C. (2002) The C-terminal domains of the RNA polymerase alpha subunits: contact site with Fis and localization during co-activation with CRP at the *Escherichia coli* proP P2 promoter. *J. Mol. Biol.*, **316**, 517–529.
57. Richet,E. (2000) Synergistic transcription activation: a dual role for CRP in the activation of an *Escherichia coli* promoter depending on MalT and CRP. *EMBO J.*, **19**, 5222–5232.
58. Belyaeva,T.A., Wade,J.T., Webster,C.L., Howard,V.J., Thomas,M.S., Hyde,E.I. and Busby,S.J. (2000) Transcription activation at the *Escherichia coli* melAB promoter: the role of MelR and the cyclic AMP receptor protein. *Mol. Microbiol.*, **36**, 211–222.
59. Wade,J.T., Belyaeva,T.A., Hyde,E.I. and Busby,S.J. (2001) A simple mechanism for co-dependence on two activators at an *Escherichia coli* promoter. *EMBO J.*, **20**, 7160–7167.
60. Chahla,M., Wooll,J., Laue,T.M., Nguyen,N. and Senear,D.F. (2003) Role of protein–protein bridging interactions on cooperative assembly of DNA-bound CRP–CytR–CRP complex and regulation of the *Escherichia coli* CytR regulon. *Biochemistry*, **42**, 3812–3825.
61. Busby,S. and Ebright,R.H. (1999) Transcription activation by catabolite activator protein (CAP). *J. Mol. Biol.*, **293**, 199–213.
62. Redfield,R.J. (1993) Genes for breakfast: the have-your-cake-and-eat-it-too of bacterial transformation. *J. Hered.*, **84**, 400–404.
63. MacFadyen,L.P., Chen,D., Vo,H.C., Liao,D., Sinotte,R. and Redfield,R.J. (2001) Competence development by *Haemophilus influenzae* is regulated by the availability of nucleic acid precursors. *Mol. Microbiol.*, **40**, 700–707.
64. Michel,B., Grompone,G., Flores,M.J. and Bidnenko,V. (2004) Multiple pathways process stalled replication forks. *Proc. Natl Acad. Sci. USA*, **101**, 12783–12788.
65. Cadman,C.J. and McGlynn,P. (2004) PriA helicase and SSB interact physically and functionally. *Nucleic Acids Res.*, **32**, 6378–6387.
66. Saveson,C.J. and Lovett,S.T. (1999) Tandem repeat recombination induced by replication fork defects in *Escherichia coli* requires a novel factor, RadC. *Genetics*, **152**, 5–13.
67. Nakanishi,A., Oshida,T., Matsushita,T., Imajoh-Ohmi,S. and Ohnuki,T. (1998) Identification of DNA gyrase inhibitor (GyrI) in *Escherichia coli*. *J. Biol. Chem.*, **273**, 1933–1938.
68. Chatterji,M. and Nagaraja,V. (2002) GyrI: a counter-defensive strategy against proteinaceous inhibitors of DNA gyrase. *EMBO Rep.*, **3**, 261–267.
69. Berge,M., Mortier-Barriere,I., Martin,B. and Claverys,J.P. (2003) Transformation of *Streptococcus pneumoniae* relies on DprA- and RecA-dependent protection of incoming DNA single strands. *Mol. Microbiol.*, **50**, 527–536.
70. Karudapuram,S., Zhao,X. and Barcak,G.J. (1995) DNA sequence and characterization of *Haemophilus influenzae* dprA+, a gene required for chromosomal but not plasmid DNA transformation. *J. Bacteriol.*, **177**, 3235–3240.
71. Gwinn,M.L., Ramanathan,R., Smith,H.O. and Tomb,J.F. (1998) A new transformation-deficient mutant of *Haemophilus influenzae* Rd with normal DNA uptake. *J. Bacteriol.*, **180**, 746–748.
72. Quillardet,P., Rouffaud,M.A. and Bouige,P. (2003) DNA array analysis of gene expression in response to UV irradiation in *Escherichia coli*. *Res. Microbiol.*, **154**, 559–572.
73. Lethem,M.I., James,S.L., Marriott,C. and Burke,J.F. (1990) The origin of DNA associated with mucus glycoproteins in cystic fibrosis sputum. *Eur. Respir. J.*, **3**, 19–23.
74. Lorenz,M.G. and Wackernagel,W. (1994) Bacterial gene transfer by natural genetic transformation in the environment. *Microbiol. Rev.*, **58**, 563–602.
75. Dell'Anno,A. and Danovaro,R. (2005) Extracellular DNA plays a key role in deep-sea ecosystem functioning. *Science*, **309**, 2179.
76. Brown,C.T. and Callan,C.G., Jr (2004) Evolutionary comparisons suggest many novel cAMP response protein binding sites in *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **101**, 2404–2409.
77. Gonzalez,A.D., Espinosa,V., Vasconcelos,A.T., Perez-Rueda,E. and Collado-Vides,J. (2005) TRACTOR_DB: a database of regulatory networks in gamma-proteobacterial genomes. *Nucleic Acids Res.*, **33**, D98–D102.
78. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
79. Gralla,J.D. and Collado-Vides,J. (1996) Organization and function of transcription regulatory elements. In Neidhardt,F.N. (ed.), *Escherichia coli and Salmonella typhimurium*, Washington, D.C., Vol. II, pp. 1232–1245.