CRP AND SXY REGULATE COMPETENCE GENE PROMOTERS IN HAEMOPHILUS INFLUENZAE

by

ANDREW DAFYDD SKYE CAMERON

B.Sc., Malaspina University-College, 2000

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

Microbiology and Immunology

THE UNIVERSITY OF BRITISH COLUMBIA

April 2007

ABSTRACT

Many bacteria have the ability to bind and take up DNA from their environment through a process called natural competence. Competent species are widely distributed across the prokaryotic phylogenetic tree and inhabit disparate niches. Most of these species are thought to tightly regulate DNA uptake, suggesting that it is an important physiological response to conditions that can arise in diverse environments. To better understand some of the signals and mechanisms that regulate competence development, I have carried out molecular studies in *Haemophilus influenzae* and other γ-proteobacteria.

In *H. influenzae*, transcription of natural competence genes depends on two proteins, CRP and Sxy. In *Escherichia coli*, CRP is a well-characterized transcription factor that stimulates gene expression in response to sugar and energy starvation; genetic studies have shown that CRP plays a similar role in *H. influenzae*. Although CRP preferentially binds to DNA sites with the consensus sequence TGTGA, results presented here demonstrate that CRP targets unusual "CRP-S" sites (sequence TGCGA) in competence gene promoters. Transcription initiation at CRP-S promoters absolutely requires Sxy, unlike other CRP-regulated promoters in the cell. Results from promoter mutagenesis and *in vitro* binding assays support a model in which CRP cannot bind CRP-S sites unless assisted by Sxy. Bioinformatic analysis identified competence genes, Sxy orthologs, and CRP-S sites in three γ-proteobacteria families (*Enterobacteriaceae*, *Pasteurellaceae*, and *Vibrionaceae*), suggesting that many bacteria use CRP and Sxy to regulate competence.

Studies in *H. influenzae* identified an extensive secondary structure in *sxy* mRNA that blocks translation. Culturing cells in starvation medium improved Sxy translation while independently stimulating CRP activity at the *sxy* promoter; however, the addition of nucleotides to starvation medium blocked Sxy translation. Thus, *sxy* mRNA secondary structure is responsive to conditions where exogenous DNA can be used as a source of nucleotides, and transcription of *sxy* is simultaneously enhanced if CRP signals that energy supplies are limited. In conclusion, nutritional signals transduced by CRP and Sxy are integrated by CRP-S sites in competence gene promoters.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CO-AUTHORSHIP STATEMENT

This manuscript-based thesis includes experiments and writing contributed by co-authors.

Chapter 1 (Introduction) and Chapter 6 (Discussion) were written by A. Cameron.

Chapter 2: R. Redfield and A. Cameron conceived of the research; RR designed the microarray experiments with input from AC. RR and Q. Qian conducted microarray experiments (Figures 2.1 and 2.2), RR analyzed microarray data presented in Figure 2.3. AC analyzed microarray data presented in Figure 2.6 and Appendix 1. AC conducted CRP-binding site and genome-wide sequence analysis (Figures 2.3 and 2.4), and protein purification and bandshift analysis (Figure 2.5). J. Hinds, T.R. Ali, J.S. Kroll, and P.R. Langford developed the *H. influenzae* whole-genome microarrays; JSK and PFL contributed to writing. RR wrote 85% and AC wrote 15% of the manuscript.

Chapter 3: AC conceived and designed the research with input and guidance from RR. AC conducted the experiments and analyzed data. AC wrote 85% and RR wrote 15% of the manuscript.

Chapter 4: AC conceived and designed the research with input and guidance from RR. AC conducted the experiments and analyzed data. AC wrote the manuscript, RR edited the manuscript.

Chapter 5: All authors (AC, Milica Volar, Laura Bannister, and RR) contributed to research conception and experimental design. AC conducted transformation assays (Figures 5.1 and 5.3), antibody synthesis and protein quantification (Figures 5.3, 5.6, and 5.8), quantification of transcript abundance (Figures 5.8 and 5.9), and *in silico* RNA folding analysis. MV conducted nuclease RNA mapping experiments; MV and AC quantified and analyzed RNase data (Figure 5.5). LB generated site-directed mutants and *lacZ*-fusion strains, and quantified *lacZ* expression (Figure 5.7); RR generated and screened hypercompetence mutants. AC wrote 60% and RR wrote 40% of the manuscript.

CHAPTER ONE

General Introduction

Natural competence is the process by which many species of bacteria bind and internalize extracellular DNA. Imported DNA is degraded unless it recombines with the chromosome; recombination can result in natural transformation if the foreign DNA permanently changes the host genotype. In most species, the machinery required for DNA transport and processing is tightly regulated at the gene level (1). This thesis examines the molecular mechanisms regulating natural competence genes in the model bacterium *Haemophilus influenzae*.

THE PHYSIOLOGICAL IMPORTANCE OF NATURAL COMPETENCE

Natural competence is widespread across the prokaryotic phylogenetic tree (2). DNA uptake by such a wide diversity of organisms inhabiting very disparate environments indicates that competence is not a niche-specific physiological adaptation. Because importing foreign DNA has multiple potential benefits, as described below, controversy surrounds the selective pressures that led to the evolution of competence. Two models are favoured to explain why cells become competent: 1) exogenous DNA is a source of novel genes, and 2) exogenous DNA is a source of nutrients (3-6).

Horizontal gene transfer (HGT) within and across species boundaries can confer adaptive benefits. This has resulted in the popular belief that prokaryotes evolved mechanisms specifically to facilitate HGT (6). However, two mechanisms of HGT are byproducts of genetic parasitism by plasmids (conjugation) or phage (transduction) and therefore only natural competence may have evolved to promote genetic diversity through HGT. Unfortunately, this model does not make clear predictions about the environmental or physiological stimuli that are expected to trigger competence. Natural competence for transformation would be useful if bacteria could sense the fitness landscape and subsequently acquire genes from fitter neighbours, however no such mechanism is known to exist.

The 'DNA as food' model posits that reuse of nucleotides from exogenous DNA provides an immediate selective advantage to any competent cell. This hypothesis predicts that competence mechanisms are controlled by factors that respond to metabolic starvation. Because regulatory

mechanisms evolve by selection for the adaptive expression of the traits they control, studying the regulation of competence genes will provide insight into the utility of DNA uptake.

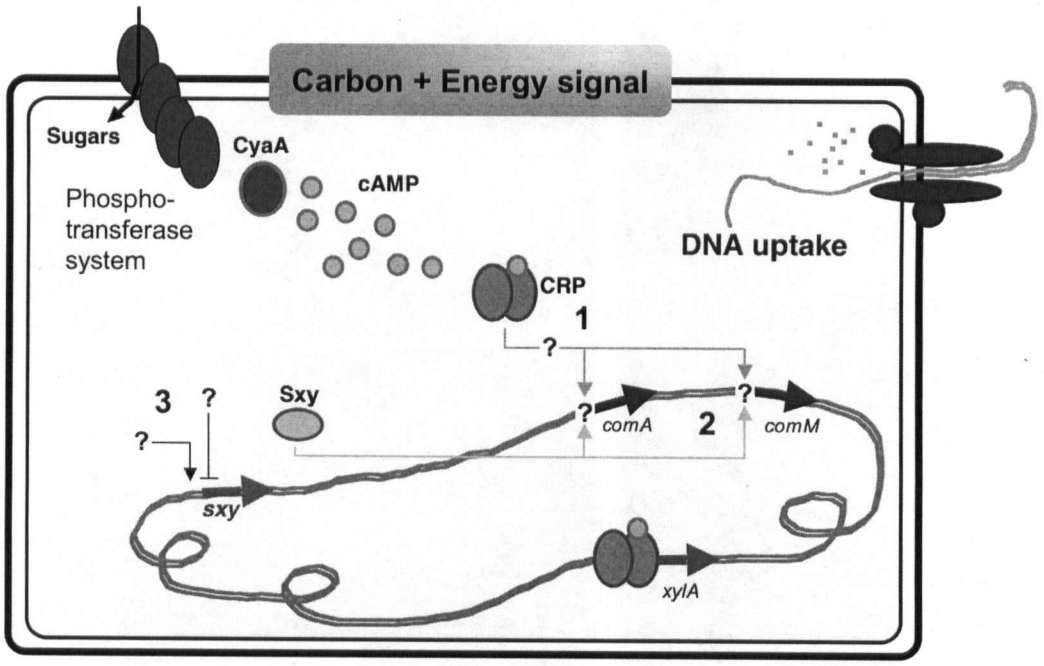
REGULATION OF NATURAL COMPETENCE

Natural competence and transformation have been best studied in *H. influenzae*, *Neisseria gonorrhea*, *Bacillus subtilis*, and *Streptococcus pneumoniae* (1). *N. gonorrhea* is constitutively competent in laboratory culture, but it is unknown whether this reflects a natural state (1). The other three species express competence genes only during specific culture conditions (*H. influenzae*) or in response to high concentrations of quorum sensing molecules (*B. subtilis* and *S. pneumoniae*). Because *H. influenzae* can be easily induced to high levels of competence, it has become a preferred species in which to study Gram-negative DNA uptake mechanisms and their regulation.

H. influenzae as a model organism for studying competence

H. influenzae is a γ-proteobacterium in the family *Pasteurellaceae*, the sister group to *E. coli*'s *Enterobacteriaceae*. The *Pasteurellaceae* are generally commensal or pathogenic inhabitants of mammalian and avian respiratory mucosa. Adaptation to obligate host niches has generated small (~2 Mbp), A+T rich (~60%) genomes with limited metabolic capabilities (7). Due in part to these genome characteristics, *H. influenzae* was the first organism to have its complete genome sequenced (8).

Though many genes are required for regulation of competence induction in *H. influenzae*, two distinct signals can be distinguished. One signal is transduced via the carbon and energy-starvation response mediated by cyclic AMP and its receptor protein (CRP); the other signal is carried by the Sxy protein (9-14). CRP and Sxy are hypothesized to meet at specialized sequences in competence gene promoters called competence regulatory elements (CRE), and then stimulate transcription. Figure 1.1 outlines three major gaps in our understanding which this thesis seeks to fill. The key players are introduced below.



Haemophilus influenzae

Figure 1.1 Regulation of competence in *H. influenzae*
This simple model of a *H. influenzae* cell illustrates the cAMP and CRP-mediated sugar-starvation response that induces genes such as *xylA*. CRP and Sxy induce expression of the DNA uptake machinery, however the regulatory mechanism(s) is unknown. Numbers in the figure highlight outstanding questions: (**1**) Does CRP directly induce competence genes? (**2**) How do competence genes integrate inducing signals? (**3**) How is *sxy* expression regulated?

Cyclic AMP signaling by the phosphotransferase system

The signaling molecule 3',5' cyclic adenosine monophosphate (cAMP) is widely used in bacteria (15). Cyclic AMP was first identified in *E. coli* and was found to regulate the diauxic switch from growth using glucose to use of less-preferred sugars (16, 17). It is now understood that the phosphotransferase system (PTS) in *E. coli* and *H. influenzae* regulates cAMP levels in response to sugar availability in the environment (11, 18). During transport across the cytoplasmic membrane, the PTS transfers a phosphate molecule from phosphoenolpyruvate to an incoming sugar. When preferred sugars are not available, phosphates accumulate on PTS proteins and trigger adenylate cyclase (CyaA) to convert ATP to cAMP (18).

H. influenzae requires *cyaA* and cAMP for competence development (9). Cells are not competent during exponential growth when intracellular cAMP concentrations are low, but as

growth slows during late exponential growth and early stationary phase, cAMP levels rise and around 1% of cells become competent (12, 14). All cells become competent ("maximal competence") when they are transferred from exponential phase to starvation medium called MIV ("M-4"); these severe starvation conditions are thought to result in maximal cAMP synthesis (14, 19).

The cAMP receptor protein

The cyclic AMP receptor protein, CRP (also called catabolite activator protein, CAP) is essential for competence development and fermentation of non-PTS sugars in *H. influenzae* (20, 21). *H. influenzae* CRP has not been well studied, but it shares 78% sequence identity with its well-characterized *E. coli* homolog. Complementation of an *H. influenzae crp* null mutant by *E. coli crp* indicates that these genes are functionally identical or very similar (20).

E. coli CRP was the first transcription factor to have its structure solved, and this has been further refined by extensive studies of CRP in complex with cAMP, DNA, and RNA polymerase (RNAP) (22-29). The larger N-terminus contains a cAMP-binding pocket and a dimerization domain. Each CRP dimer can be activated by one cAMP molecule, resulting in a conformational change that exposes each monomer's C-terminal helix-turn-helix DNA-binding domain (30). CRP dimers bind specifically to 22 base pair (bp) sequences located at or near gene promoters and recruits RNAP to initiate transcription (reviewed in (31)). Throughout this thesis, CRP dimers will be referred to simply as "CRP" for simplicity.

The *sxy* gene

The *sxy* gene encodes a positive regulator of competence. It was first discovered as the site of a gain-of-function mutation (*sxy-1*) that cause elevated competence (hypercompetence) in non-competence inducing conditions (32). Deletion of *sxy* abolishes competence development, but no other phenotype has been associated with this mutant (10). Sxy shares no homology with any characterized protein or protein domains and its mode of action has remained speculative.

Without the *crp* or *sxy* gene products, essential competence genes are not expressed, so CRP and Sxy must act early in the competence-inducing cascade (13, 14). Several results suggest that CRP and Sxy work in concert to promote competence. The addition of 1 mM cAMP to culture medium, which is expected to elevate CRP activity, leads to an increase in competence but does not result in maximal competence (9). On the other hand, the mutation of a specific base pair in

the *sxy* coding region, the *sxy-1* mutation, results in the overexpression of *sxy* and increased, but not maximal, competence in rich medium (13). Thus, neither CRP nor Sxy alone is able to induce maximal competence, yet if cAMP is added to *sxy-1* cells in rich medium, MIV-induced competence levels are achieved (10, 32).

The competence regulatory element

The competence regulatory element (CRE) was identified as a 22 bp palindromic nucleotide sequence in the promoter regions of some essential competence genes (33, 34). The loss of competence gene expression as a result of transposon mutagenesis at or near CRE sites led investigators to suspect that these sequences are essential for competence gene expression ((35); C. Ma, personal communication). In addition, the presence of CREs in the promoter regions of multiple competence genes suggested that this sequence is important for gene regulation (33, 36). Initially, CREs were thought to be Sxy-binding sites, but Macfadyen (34) showed that CREs resemble CRP-binding sites and proposed a model in which competence gene expression is activated when CRP binds CREs. Moreover, the absence of a recognizable DNA-binding domain in the Sxy protein led Macfadyen (34) to suggest that Sxy interacts with CRP, increasing CRP's affinity for CRE sites.

Transcription activation: RNA polymerase recruitment by transcription factors

All bacterial genes are regulated through control of the rate at which RNA polymerase (RNAP) initiates transcription (37). RNAP is composed of five protein subunits (α, α, β, β', and σ), three of which (α, α, and σ) recognize promoter DNA. RNAP alone can efficiently initiate transcription at some promoters; however, the stimulation of most promoters is fine-tuned by transcription factors that convert physiological and physiochemical signals into gene expression (reviewed in (37)). One mechanism for transcriptional activation is to recruit RNAP to promoters for which the polymerase has low intrinsic affinity (38); CRP is the classic example of a transcription factor that recruits RNAP (29). Sxy may operate in a similar capacity by recruiting RNAP to competence gene promoters in response to an as yet unidentified environmental or cellular signal.

TRANSCRIPTIONAL REGULATORY NETWORKS

Because a bacterium must at all times satisfy multiple metabolic requirements, it needs to continuously balance its internal functions while exploiting a potentially ever-changing external environment. Consequently, bacteria exhibit complex yet well-integrated responses to a wide variety of environmental, physiological, and physiochemical stimuli (39). Stimuli are transduced and interpreted through networks of sensory proteins and transcription factors; it is impressive that most species use a mere 50 to 500 transcription factors to respond and adapt to environmental changes (40). Not surprisingly, bacteria that inhabit very stable niches (such as living within the cells of symbiotic hosts) have a small number of transcription factors, whereas bacteria in more complex environments employ a much larger number of regulators (41). This relationship has been shown to scale as a power-law in which the number of transcription factors doubles twice as fast as does the total number of genes in a genome (40, 42), indicating that large bacterial genomes employ disproportionately more complex regulatory networks.

Transcriptional regulatory networks within a genome also follow a power-law distribution. Thus, a small number of transcription factors (called global regulators) regulate a large number of genes (43). Global regulators are often described as well-connected nodes within the regulatory network (42), and it has been observed that their high degree of pleiotropy correlates with a decrease in DNA-site specificity (44). At the other end of the spectrum, most of a cell's transcription factors target highly specific DNA sites in one or a few promoters.

The term "regulon" is used to describe all genes whose transcription is regulated by the same transcription factor. However, all transcriptional units within a regulon may not be coordinately expressed. This is because each promoter can integrate signals from multiple regulons and is also subject to a variety of other factors, including the binding of RNAP σ subunits, DNA methylation states, and local DNA topology mediated by DNA-bending proteins (37, 45-48). In other words, two promoters can share a common transcription factor, but each promoter's expression may be conditional on a separate, unconnected signal.

A variety of interactions have been observed at promoters targeted by multiple transcription factors (reviewed in (46)). These interactions may be cooperative or antagonistic and can range from direct protein-protein contacts to indirect interactions mediated by changes in DNA topology. For example, CRP binds cooperatively with MelR to stimulate transcription of the *melAB* promoter (49). At the more complex *nrf* promoter, IHF and NarP/NarL compete for

overlapping binding sites; IHF changes DNA topology such that FNR cannot stimulate transcription whereas NarP/NarL reverse this repression (50). This heterogeneity of promoter architectures contributes to the complexity of bacterial transcriptional networks.

THESIS OBJECTIVES

This thesis seeks to uncover the molecular mechanism(s) responsible for competence gene induction. Five general hypotheses are addressed: (1) Competence genes are united in a Sxy and CRP-dependent regulon, (2) CRP binds specifically to CREs, (3) Sxy helps CRP bind DNA and activate transcription at CRE promoters, (4) The *sxy* gene is upregulated in competence-inducing conditions, (5) *H. influenzae*'s close relatives also use CRP, Sxy, and CREs to regulate competence genes.

OVERVIEW OF THE CHAPTERS IN THIS THESIS

This thesis is composed of four manuscripts; each describes research results and analyses generated by myself and collaborators.

Chapter 2 describes the use of whole-genome microarrays to follow changes in gene expression during competence development in wildtype *H. influenzae* cells. Microarrays were also used to characterize dependence of MIV-induced transcription on CRP and Sxy. This analysis provided evidence for the existence of a competence regulon, characterized by a promoter-associated 22 bp competence regulatory element (CRE) closely related to the cAMP receptor protein (CRP) binding consensus. This CRE regulon contains 25 genes in 13 transcription units, only about half of which have been previously associated with competence. Bandshift assays confirmed that CRE sequences are a new class of CRP-binding site. The essential competence gene *sxy* is induced early in competence development and is required for MIV-induced transcription of CRE-regulon genes but not other CRP-regulated genes, suggesting that Sxy may act as an accessory factor directing CRP to CRE sites.

Chapter 3 introduces the name "CRP-S" to replace the more ambiguous name "CRE". CRP-S sites are defined by the core sequence TGCGA, distinguishing them from canonical (CRP-N) sites with the core TGTGA. First we report that all γ-proteobacteria encode orthologs of *H. influenzae*'s competence genes, whereas sxy orthologs are found only in the *Enterobacteriaceae*, *Pasteurellaceae*, and *Vibrionaceae*. Phylogenetic footprinting identified CRP-S and CRP-N sites in *Enterobacteriaceae*, *Pasteurellaceae*, and *Vibrionaceae* genomes

that we analyzed. Bandshift experiments confirmed that *E. coli* CRP-S sequences are CRP binding sites, and mRNA analysis showed that they require CRP, cAMP, and Sxy for gene induction.

Chapter 4 describes a detailed analysis of CRP binding to CRP-S and CRP-N promoters. We found that *E. coli* CRP has a higher non-specific affinity for DNA than does *H. influenzae* CRP. *H. influenzae* CRP was found to be very discriminating in terms of which sites it will bind; for example, it cannot bind a CRP-S site *in vitro* unless the TGCGA sequence is converted to TGTGA. Further results implicated Sxy in facilitating CRP binding to DNA and in helping recruit RNAP, possibly by mediating contacts between RNAP and UP elements in CRP-S promoters.

Chapter 5 describes mutations in *sxy* that elevate Sxy protein levels by 7-25 fold, which results in hypercompetence. *In vitro* nuclease analysis confirmed the existence of an extensive 2° structure at the 5' end of *sxy* mRNA that sequesters the ribosome binding site. Hypercompetence mutations were found to reduce base pairing in this structure, causing a global destabilization that exposes 5' mRNA for ribosome binding. Conversely, mutations engineered to add base pairs strengthen mRNA folding, reduce translation, and greatly reduce competence. Starvation medium is shown to improve Sxy translation while independently stimulating CRP activity at the *sxy* promoter.

REFERENCES

1. Solomon,J.M. and Grossman,A.D. (1996) Who's competent and when: regulation of natural genetic competence in bacteria. *Trends Genet.*, **12**, 150-155.
2. Lorenz,M.G. and Wackernagel,W. (1994) Bacterial gene transfer by natural genetic transformation in the environment. *Microbiol. Rev.*, **58**, 563-602.
3. Redfield,R.J. (1993) Genes for breakfast: the have-your-cake-and-eat-it-too of bacterial transformation. *J Hered,* **84**, 400-404.
4. Suerbaum,S., Smith,J.M., Bapumiá,K., Morelli,G., Smith,N.H., Kunstmann,E., Dyrek,I. and Achtman,M. (1998) Free recombination within Helicobacter pylori. *Proc Natl Acad Sci U S A,* **95**, 12619-12624.
5. Dubnau,D. (1999) DNA uptake in bacteria. *Annu. Rev. Microbiol.*, **53**, 217-244.
6. Redfield,R.J. (2001) Do bacteria have sex? *Nat Rev Genet,* **2**, 634-639.
7. Redfield,R.J., Findlay,W.A., Bosse,J., Kroll,J.S., Cameron,A.D. and Nash,J.H. (2006) Evolution of competence and DNA uptake specificity in the Pasteurellaceae. *BMC Evol Biol,* **6**, 82.
8. Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.F., Dougherty,B.A., Merrick,J.M. and et,a. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science,* **269**, 496-512.
9. Dorocicz,I.R., Williams,P.M. and Redfield,R.J. (1993) The Haemophilus influenzae adenylate cyclase gene: cloning, sequence, and essential role in competence. *J. Bacteriol.*, **175**, 7142-7149.
10. 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.
11. Macfadyen,L.P., Dorocicz,I.R., Reizer,J., Saier,M.H.J. and Redfield,R.J. (1996) Regulation of competence development and sugar utilization in Haemophilus influenzae Rd by a phosphoenolpyruvate:fructose phosphotransferase system. *Mol. Microbiol.*, **21**, 941-952.
12. Macfadyen,L.P., Ma,C. and Redfield,R.J. (1998) A 3',5' cyclic AMP (cAMP) phosphodiesterase modulates cAMP levels and optimizes competence in Haemophilus influenzae Rd. *J. Bacteriol.*, **180**, 4401-4405.
13. Bannister,L.A. (1999) An RNA secondary structure regulates sxy expession and competence development in Haemophilus influenzae. PhD Thesis, University of British Columbia.
14. Macfadyen,L.P. (1999) PTS regulation of competence in Haemophilus influenzae. PhD Thesis, University of British Columbia.
15. Botsford,J.L. and Harman,J.G. (1992) Cyclic AMP in prokaryotes. *Microbiol. Rev.*, **56**, 100-122.
16. Makman,R.S. and Sutherland,E.W. (1965) Adenosine 3',5'-phosphate in Escherichia coli. *J. Biol. Chem.*, **240**, 1309-1314.

17. Pastan,I. and Perlman,R. (1970) Cyclic adenosine monophosphate in bacteria. *Science,* **169**, 339-344.

18. Postma,P.W., Lengeler,J.W. and Jacobson,G.R. (1996) Phosphoenolpyruvate:carbohydrate phosphotransferase system. In Neidhardt,F.N., et al. (ed.), *Escherichia coli and Salmonella typhimurium,* Washington, D.C, Vol. II, pp. 1149-1174.

19. Herriott,R.M., Meyer,E.M. and Vogt,M. (1970) Defined nongrowth media for stage II development of competence in Haemophilus influenzae. *J. Bacteriol.,* **101**, 517-524.

20. Chandler,M.S. (1992) The gene encoding cAMP receptor protein is required for competence development in Haemophilus influenzae Rd. *Proc Natl Acad Sci U S A,* **89**, 1626-1630.

21. Macfadyen,L.P. and Redfield,R.J. (1996) Life in mucus: sugar metabolism in Haemophilus influenzae. *Res Microbiol,* **147**, 541-551.

22. 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.

23. McKay,D.B., Weber,I.T. and Steitz,T.A. (1982) Structure of catabolite gene activator protein at 2.9-A resolution. Incorporation of amino acid sequence and interactions with cyclic AMP. *J. Biol. Chem.,* **257**, 9518-9524.

24. Weber,I.T. and Steitz,T.A. (1987) Structure of a complex of catabolite gene activator protein and cyclic AMP refined at 2.5 A resolution. *J. Mol. Biol.,* **198**, 311-326.

25. 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.

26. 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.

27. Passner,J.M. and Steitz,T.A. (1997) The structure of a CAP-DNA complex having two cAMP molecules bound to each monomer. *Proc Natl Acad Sci U S A,* **94**, 2843-2847.

28. Passner,J.M., Schultz,S.C. and Steitz,T.A. (2000) Modeling the cAMP-induced allosteric transition using the crystal structure of CAP-cAMP at 2.1 A resolution. *J. Mol. Biol.,* **304**, 847-859.

29. Benoff,B., Yang,H., Lawson,C.L., Parkinson,G., Liu,J., Blatter,E., Ebright,Y.W., Berman,H.M. and Ebright,R.H. (2002) Structural basis of transcription activation: the CAP-alpha CTD-DNA complex. *Science,* **297**, 1562-1566.

30. Harman,J.G. (2001) Allosteric regulation of the cAMP receptor protein. *Biochim. Biophys. Acta,* **1547**, 1-17.

31. Busby,S. and Ebright,R.H. (1999) Transcription activation by catabolite activator protein (CAP). *J. Mol. Biol.,* **293**, 199-213.

32. Redfield,R.J. (1991) sxy-1, a Haemophilus influenzae mutation causing greatly enhanced spontaneous competence. *J. Bacteriol.,* **173**, 5612-5618.

33. Karudapuram,S. and Barcak,G.J. (1997) The Haemophilus influenzae dprABC genes constitute a competence-inducible operon that requires the product of the tfoX (sxy) gene for transcriptional activation. *J. Bacteriol.,* **179**, 4815-4820.

34. Macfadyen,L.P. (2000) Regulation of competence development in Haemophilus influenzae. *J Theor Biol,* **207**, 349-359.

35. 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.

36. 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.

37. Browning,D.F. and Busby,S.J. (2004) The regulation of bacterial transcription initiation. *Nat Rev Microbiol,* **2**, 57-65.

38. Ptashne,M. and Gann,A. (1997) Transcriptional activation by recruitment. *Nature,* **386**, 569-577.

39. Cases,I. and de Lorenzo,V. (2005) Promoters in the environment: transcriptional regulation in its natural context. *Nat Rev Microbiol,* **3**, 105-118.

40. van Nimwegen,E. (2003) Scaling laws in the functional content of genomes. *Trends Genet.,* **19**, 479-484.

41. Moran,N.A., Dunbar,H.E. and Wilcox,J.L. (2005) Regulation of transcription in a reduced bacterial genome: nutrient-provisioning genes of the obligate symbiont Buchnera aphidicola. *J. Bacteriol.,* **187**, 4229-4237.

42. Madan Babu,M., Teichmann,S.A. and Aravind,L. (2006) Evolutionary dynamics of prokaryotic transcriptional regulatory networks. *J. Mol. Biol.,* **358**, 614-633.

43. Martinez-Antonio,A. and Collado-Vides,J. (2003) Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr Opin Microbiol,* **6**, 482-489.

44. Sengupta,A.M., Djordjevic,M. and Shraiman,B.I. (2002) Specificity and robustness in transcription control networks. *Proc Natl Acad Sci U S A,* **99**, 2072-2077.

45. Low,D.A., Weyand,N.J. and Mahan,M.J. (2001) Roles of DNA adenine methylation in regulating bacterial gene expression and virulence. *Infect. Immun.,* **69**, 7197-7204.

46. 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.

47. McLeod,S.M. and Johnson,R.C. (2001) Control of transcription by nucleoid proteins. *Curr Opin Microbiol,* **4**, 152-159.

48. Hernday,A.D., Braaten,B.A. and Low,D.A. (2003) The mechanism by which DNA adenine methylase and PapI activate the pap epigenetic switch. *Mol Cell,* **12**, 947-957.

49. 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.

50. Browning,D.F., Cole,J.A. and Busby,S.J. (2000) Suppression of FNR-dependent transcription activation at the Escherichia coli nir promoter by Fis, IHF and H-NS: modulation of transcription initiation by a complex nucleo-protein assembly. *Mol. Microbiol.,* **37**, 1258-1269.

CHAPTER TWO

A novel CRP-dependent regulon controls expression of competence genes in *Haemophilus influenzae*[1]

INTRODUCTION

In *Haemophilus influenzae*, competence genes have been identified primarily by screens for mutants defective in transformation, and have been tentatively assigned regulatory or mechanistic roles based on mutant phenotypes and homology to competence genes in other bacteria. Most of the known *H. influenzae* competence genes are thought to encode proteins with direct roles in DNA uptake or in assembly of the uptake machinery. These include proteins acting at the outer membrane (*comE*, *pilA*), in the periplasm or at the inner membrane (*comC*, *comF* (=*com101A*), *rec-2*), and cytoplasmically (*comA*, *dprA*, *comM*) (1-6). There have been several non-exhaustive screens for transformation-defective mutants (7-11); each has identified some new candidate genes but missed known genes, suggesting that some regulatory and DNA uptake genes may not yet be identified.

H. influenzae is the Gram-negative bacterium whose competence regulation is best understood, and microarray analysis of regulatory mutants allowed us to investigate this regulation. Two regulatory proteins, CRP and Sxy, are required to activate transcription of *H. influenzae* competence genes. CRP is the cAMP regulatory protein, best characterized in *Escherichia coli*; it activates transcription of many genes including the carbon-energy regulon when rising cAMP levels signal that preferred sugars are unavailable. Like *E. coli*, *H. influenzae* has a phosphotransferase system that regulates cAMP levels and thus gene activation by CRP (15-18). In *H. influenzae* cAMP and CRP also regulate competence; *crp* and *cya* (adenylate cyclase) mutants are unable to become competent (19, 20) or to induce expression of the competence gene *comA* (9). Regulation of competence by such energy-supply signals is consistent with its proposed role in nutrient acquisition. In *H. influenzae* as in *E. coli*, CRP regulates diverse genes involved in nutrient acquisition or use (21), and most genes regulated by CRP are predicted to be subject to additional function-specific regulation. Addition of cAMP to exponentially

---

[1] A version of this chapter has been published. Redfield R.J., Cameron A.D.S., 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.* **374**: 735-747

growing wild-type cells does not induce maximal competence, but only the 100-fold lower competence also seen at the onset of stationary phase (22), suggesting that competence genes may be subject to regulation by another, competence-specific regulator.

The Sxy protein may be the predicted second regulator. Strains overexpressing Sxy have elevated competence, and a *sxy* knockout mutant, like a *crp* mutant, is unable to become competent and fails to induce expression of *lacZ* fusions to the *comF* and *dprA* genes (5, 23, 24). However, we understand neither how Sxy regulates competence genes nor how Sxy itself is regulated. Although Sxy lacks the structural features typical of DNA binding proteins, it has been postulated to activate transcription by binding to DNA at the competence regulatory element sequences associated with promoters of DNA uptake genes (see below) (5). A role for mRNA secondary structure in regulating *sxy* expression is likely: point mutations weakening base pairing in a 5' stem of *sxy* mRNA dramatically increase *sxy* mRNA expression and competence; mutations strengthening the stem eliminate both (23, 25).

The hypothesis that competence regulatory element (CRE) sequences in the promoters of DNA uptake genes are responsible for competence-specific regulation of transcription was strengthened by Gwinn *et al.*'s demonstration that *comM*, a gene initially identified only by the presence of a CRE sequence in the promoter, is competence-regulated and essential for transformation (5, 6). Only two symmetrical base pairs distinguish the core CRE sequence TGCGA(N_6)TCGCA from the core CRP binding sequence TGTGA(N_6)TCACA, suggesting that CRP may bind to CRE sites and activate transcription directly, rather than regulating competence indirectly by regulating *sxy* transcription (24). Under this model, the presence of cAMP and Sxy allows CRP to bind at CRE sites and activate transcription of DNA uptake genes (26). Consistent with this, both CRP and Sxy are required for transcription of genes in the *comA-F* operon (18, 24).

We have generated microarrays containing all 1738 genes of the sequenced *H. influenzae* strain KW20, and have used them to characterize (1) the changes in expression of the 1738 *H. influenzae* genes in response to transfer from rich medium to MIV, and (2) the extent to which these changes depended on the presence of Sxy and CRP. This enabled us to show that the subset of starvation-induced genes that possess CRE sequences are also united by their requirement for both CRP and Sxy. This CRE regulon includes most of the identified

components of the DNA uptake machinery, in addition to a number of new genes not previously associated with competence.

MATERIALS AND METHODS:

Microarray slide preparation.

The *H. influenzae* whole genome microarray was based on the annotated sequence of the Rd strain (49). Primer3 software (50) was used to design primer pairs to amplify an internal sequence of each ORF. Software parameters dictated the annealing temperatures of approximately 55°C and PCR product sizes between approximately 175-600bp. BLAST analysis was used to minimize homology with other ORFs within the genome. Sizes of PCR products were checked using agarose gels. Reactions with multiple or no products were repeated at lower and higher annealing temperatures, and those which produced incorrect-sized products had their primers redesigned.

PCR products of all 1738 *H. influenzae* genes were spotted in duplicate onto poly-L-lysine-coated glass microscope slides by a MicroGridII robot (BioRobotics, UK), using the facilities of the Bacterial Microarray Group at St. George's Hospital Medical School, London. Control spots were: *H. influenzae* 5s, 16s and 23s rRNA genes; human and rat actin genes; and *E. coli lacZ* and *glpD* genes. tRNA genes were not included. Slide processing prior to hybridization has been previously described (51). Quality controls used the first and last slides of each print run.

Strains and growth conditions

KW20 is the standard *H. influenzae* Rd strain sequenced by Fleischman *et al.* (49). The MAP7, *cya*, and *sxy* knockout strains have been described (19, 20, 23, 48). Culture growth and competence protocols have been described (48, 53). MIV medium contains (all amounts in μg/ml): Arg, 21; Asp, 4032; Cys, 6; Glu, 314; Leu, 61; Lys, 35; Met, 18; Ser, 65; Tyr, 42; Ile, 33; Gly, 2.5; His, 13; Val, 35; Phe, 46; Thr, 20; Ala, 48; Pro, 50; Fumarate, 1000; Citrulline, 12; Tween-80, 200; NaCl, 4675; $MgSO_4$, 124; $CaCl_2$, 147; KH_2PO_4, 1740. Cultures used for the time courses were pregrown in sBHI at densities below 2×10^8 cfu/ml for at least two hours before the first time point was taken. Sample times are specified relative to t=0 min, when cells in sBHI at a density of 8×10^8 (OD_{600}=0.2) were transferred to MIV. Time course samples were taken from cells in sBHI at t= -70, -30, 45, 80, and 130 minutes, and from cells in MIV at t=10,

30, 60 and 100 minutes. Competence of the 100 minutes sample was confirmed by transformation to novobiocin resistance with DNA of the Nov^R strain MAP7. Samples for *cya* (four replicate experiments) and *sxy* analysis (five replicate experiments) were taken after 100 minutes of incubation in MIV.

RNA methods

Aliquots of cells (usually 2 ml) were taken from liquid cultures, pelleted (1 min at 10,000g), quick-chilled and stored frozen at -80°C. RNAs were prepared from these pellets using Qiagen RNeasy kits, and were freed of contaminating DNA with either Qiagen on-column DNase I digestion or an Ambion DNA-Free kit. RNA concentrations were determined spectrophotometrically and RNA quality was assessed by gel electrophoresis.

Microarray methods

cDNAs from signal and control RNAs were labeled with either Cy3 and Cy5 or Cy5 and Cy3 respectively in corresponding replicate experiments to limit artifacts caused by the potential differences of Cy3 or Cy5 in labeling efficiency. The labeling and hybridization procedures followed either of two protocols. Production of cDNA probes labeled with Cy3 and Cy5 and microarray hybridization used either of two protocols, one developed by the Bacterial Microarray Group for labeling with Cy3-dUTP and Cy5-dUTP (54) and the other by TIGR for amino-allyl labeling with Cy3 and Cy5 [http://www.tigr.org/tdb/microarray/protocolsTIGR.shtml]. For analysis of time-course samples, a control RNA pool containing equal amounts of RNA from all nine samples was prepared and used as competitor for each sample. This improves the quantitation of RNAs that are expressed at very low levels in some samples (13).

Analysis of microarray data

Microarray slides were scanned and intensity data was collected from the images using Imagene software (BioDiscovery).

Time course data

Data for virtual t=0 minutes samples from the time courses were created by averaging the two exponential growth samples (t=-70 and t=-30 minutes). The data were imported into GeneSpring (Silicon Genetics, v6.0) and assembled into the multi-sample 'experiments'

indicated in Figure 2.1 (*e.g.* Fig. 2.2). Datasets were normalized using GeneSpring's default 'per spot' normalization step and a modified 'per chip' normalization that restricted the measurements used in the calculation of the median to current normalized values of at least 0.01. In addition, extra background correction was applied when needed.

For cAMP supplementation and *sxy* mutant data, replicate slides were combined into a ±cAMP dataset and a ±*sxy* dataset. Each dataset was normalized using the default 'per chip' step and a refined 'per spot' step that decreased the cut off value of the control channel from 10 to 0.01 to improve the spot-detection sensitivity.

Quantitative PCR

RNAs were prepared from an independent MIV time course (–40 and 0 minutes in sBHI, 20, 60 and 100 minutes in MIV), from wild type and *sxy* knockout cells at 60 and 100 minutes in MIV, and from *cya* knockout cells at 60 and 100 minutes in MIV ± 1 mM cAMP. cDNA templates were generated using the iScript cDNA synthesis kit (BioRad). Reactions were carried out in duplicate in a 7000 SDS (Sequence Detection System) (Applied Biosystems) using the iTaqSYBR Green Supermix with Rox (BioRad) and primers designed with Primer Express 2.0 (Applied Biosystems) and on-line Net Primer for PCR products. The standard curves used five serial 5-fold dilutions of a MAP7 genomic DNA template. Relative RNA abundance measurements were calculated by normalizing derived quantity of cDNA template (*ssb* or *comF*) to that of a control (*murG*), chosen because of its strong constant expression in the microarray time courses.

Electrophoretic mobility-shift assays

Fragments containing the *mglBAC* and *comA-F* promoters (each 130 bp) were PCR-amplified from *H. influenzae* genomic DNA, purified on a 5% acrylamide gel, and internally labeled in 12.5 µl reactions containing 50 ng DNA, 6 µM of each PCR primer, 50 µM d(C,G,T)TP mix, 2 µM dATP, 0.8 µM (^{33}P) αdATP (2500 Ci/mmol), 1x Klenow buffer, and 2U Klenow enzyme. DNA and oligonucleotides were heated to 94°C for 3 minutes then placed on ice. Nucleotides, buffer, and enzyme were then added and the reaction was incubated at room temperature for 2 hours. The reaction was stopped by heating to 80°C for 20 minutes, diluted in 150 µl of TE, and stored at –20°C. CRP was purified from *E. coli* (DH5α) cells carrying the plasmid pXN15, which encodes *E. coli crp* and its native promoter (19, 55). Protein purity was assessed using

SDS-PAGE and Coomassie staining and protein concentration was measured using the BioRad DC Protein (Lowry) assay. Binding reactions (10 µl) contained 10 mM Tris HCl pH 8.0, 50 mM KCl, 0.5 mM EDTA, 10% glycerol, 250 µg/ml BSA, 100 µM cAMP, 1 mM dithiothreitol, 40 µg/ml poly(dIdC) DNA, 2.8 ng labeled DNA (400,000 CPM/ng), and purified CRP as indicated. Reactions were incubated at room temperature for 10min then loaded onto a prerun polyacrylamide gel (30:1 acrylamide:bisacrylamide, 1/5 x TBE, 2% glycerol, and 200 µM cAMP; running buffer 1/5 x TBE and 200 µM cAMP). Following electrophoresis for 2.5 hrs at 100 V, the gel was dried and exposed for 1hr to a phosphor screen. Bands were visualized using a STORM 860 scanner (Applied Biosystems).

Sequence analysis

The program RSA-tools was used to search for sequences resembling CRE sites (27). The input matrix was based on the first 9 CRE sequences in Fig. 3, using the calculations described by Macfadyen (26). Sequence motifs were identified using the programs Consensus (30), Gibbs recursive sampler (31) and Bioprospector (32). Sequence logos were generated using WebLogo (56).

Database deposition:

Fully annotated data from these arrays have been placed in the BmG@Sbase, accession no. E-BUGS-20 (http://bugs.sghms.ac.uk/E-BUGS-20) and ArrayExpress accession no. E-BUGS-20.

RESULTS

Figure 2.1 illustrates the split time course analysis used to characterize gene expression changes during competence development. Cells were sampled both after transfer to the starvation medium MIV and during growth in the rich medium sBHI. As Figure 2.1 indicates, cells reach peak competence after 100 minutes in MIV (transformation frequency (TF) with MAP7 DNA about $3x10^{-3}$), and also become moderately competent (TF about 10^{-4}) when growth in sBHI first slows; this 'late log' competence occurs before the complete cessation of growth.

The complete time course was done twice, using RNA preparations from independent cultures. For seven of the nine time points from the two replicate time-course experiments, the replicate measurements of expression levels of 89%-92% of the genes were within twofold. The exceptions were the t=10 and t=30 minutes time points for cells in MIV, which had 81%-82% of

their values within twofold. Most differences greater than twofold were due to minor differences in the timing of competence development between the replicates, rather than to random variation. For 24 of the 27 microarrays used, the transcripts of less than 2% of the 1738 genes produced 'No Data' reports. The other three were less than 4%. This indicated that the majority of the genes on the microarray slide were transcribed regardless of the various culture conditions and cell types used in this study. In addition, it also demonstrated that the current microarray methodology and analysis system were sensitive enough to detect even transcripts of low abundance.

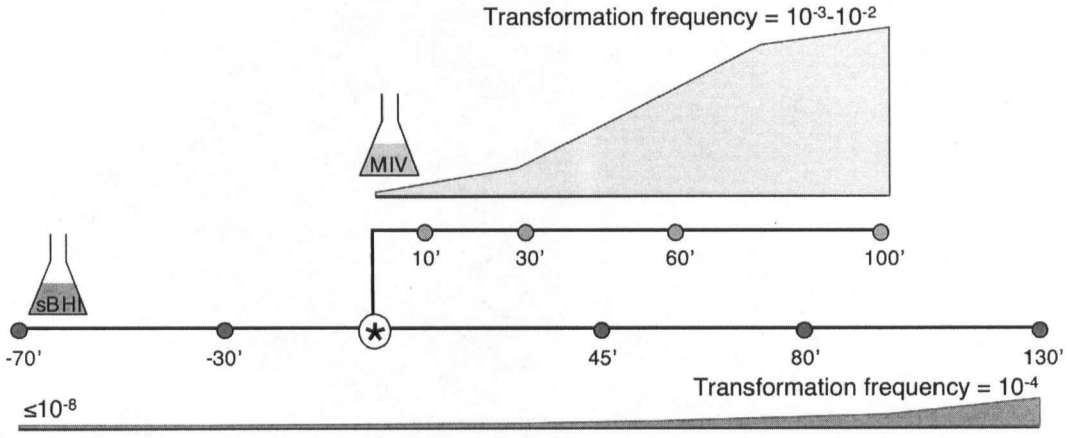


Figure 2.1 The sampling protocol for the competence time courses.
See Materials and Methods for details.

Typical microarray data from one of these time courses is shown in the panels of Figure 2.2. Figure 2.2A shows relative expression of all the 1738 genes in the *H. influenzae* genome during exponential growth (t=0) and at 10, 30, 60 and 100 minutes after transfer to MIV. 151 genes showed reproducible >4-fold increases in mRNA after transfer to MIV. Although many genes showed modest decreases in expression after transfer to MIV, only 44 decreased by at least 4-fold (lists of these genes are provided as Appendix 1). Below, we focus mainly on the genes likely to play roles in competence.

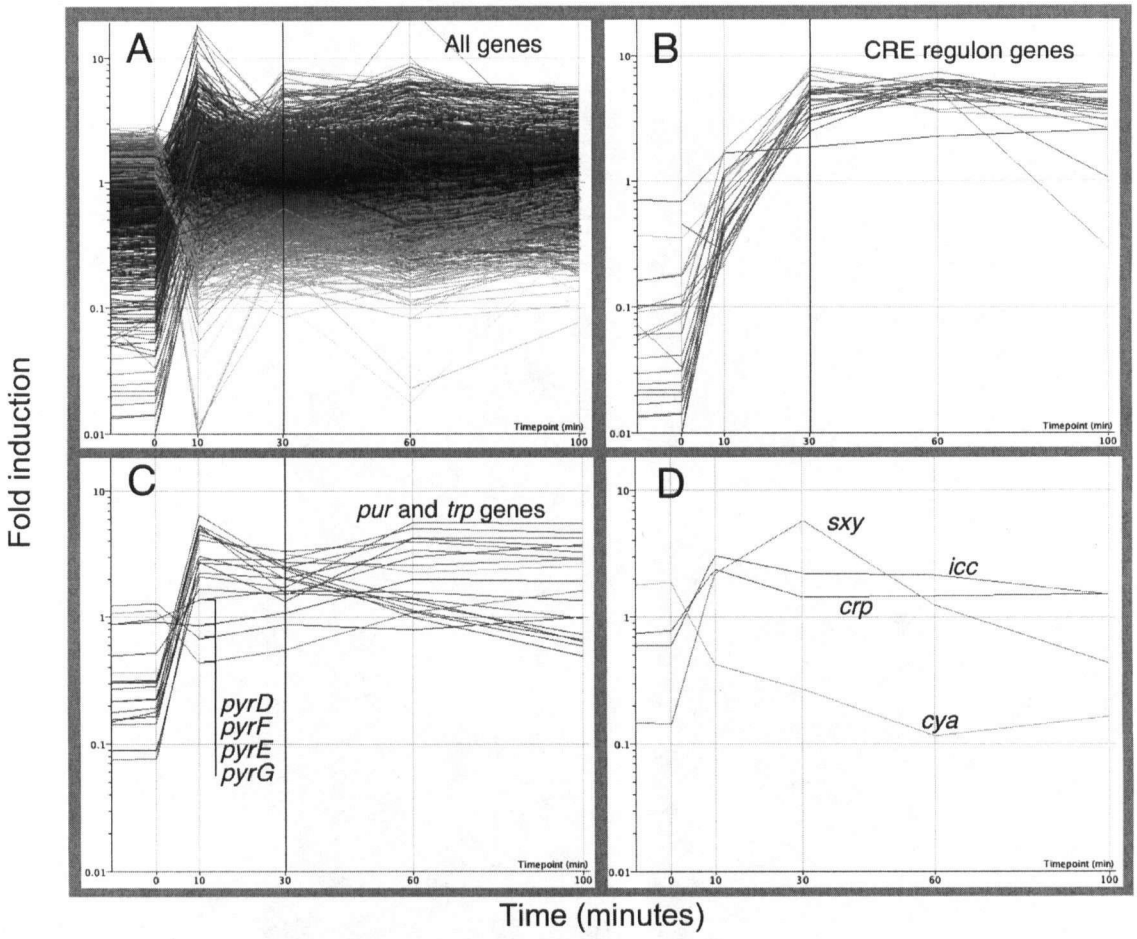


Figure 2.2 Data from microarray analysis of a competence time course.
The t=0 minutes data is the mean of t= -30 and t=-70 minutes samples. **A.** All genes on the array. **B.** All genes in the CRE regulon (listed in Figure 3). **C.** *purB*, *purC*, *purD*, *purE*, *purF*, *purH*, *purK*, *purL*, *purM*, *purN*, *trpA*, *trpB*, *trpC*, *trpD*, *trpE*, *trpG*, *pyrD*, *pyrE*, *pyrF*, *pyrG*. **D.** *sxy*, *crp*, *cya*, *icc*.

Identification of competence-induced genes

Known competence genes

We first examined starvation-induced changes in expression of the known competence genes *comA, comC, comE, comF* (all in the putative *comABCDEF* operon), *rec-2, dprA, comM,* and *pilA*; all five promoters contain previously identified CRE sequences (5). All genes except *comF* were induced strongly (45-450-fold) but more slowly than the majority of induced genes, with maximum expression usually seen at the *t*=60 minutes sample (Fig. 2.2B and Fig. 2.3). Expression levels of *comF* were low but quantitative PCR showed that it is induced about 40-fold in MIV, confirming previous reports (2, 24). As expected, *comB* and *comD* were co-induced with the rest of the *comABCDEF* operon, and *pilB*, *pilC* and *pilD* were coinduced with

pilA. However, *dprB* and *dprC* showed little induction and did not appear to be coordinately expressed with *dprA*.

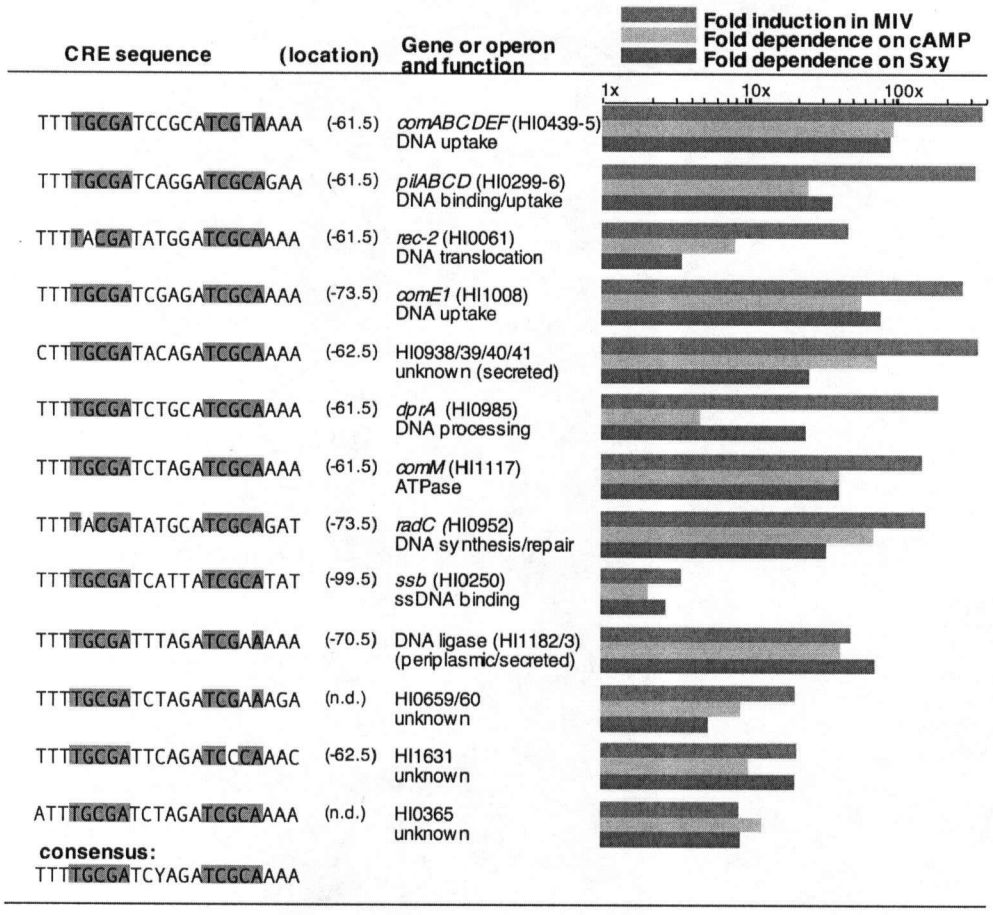


Figure 2.3 The CRE regulon.
CRE location is distance from the putative 5' end of the transcript to the center of the CRE; n.d, promoter not determined. Fold induction is the ratio of maximum expression in MIV expression level to *t*=0 expression. Fold dependence is the ratio of expression with cAMP or Sxy to expression without, after 100 min incubation in MIV.

Other CRE-regulated genes

Four uncharacterized genes had also been identified as having promoters with putative CRE sequences (4, 26). Two of these were induced with the same kinetics as the above genes: *comE1* (HI1008, 270-fold) and *ssb* (HI0250, 3.4-fold). The low but consistent induction of *ssb* was confirmed by quantitative PCR. Genes of unknown function downstream from the two other CRE sequences were also induced (HI0365, 9-fold, and HI1182, 50-fold), however these CRE sequences had originally been incorrectly assigned to the divergently transcribed genes HI0364 and HI1181 respectively. Like the known competence genes, induction of HI1008,

HI0250, HI0365 and HI1182 was relatively slow, with expression peaking at 30-60 minutes after transfer (Fig. 2.2B).

Several complementary strategies were used to search for additional genes in the CRE regulon, and to exclude others from it. First, the MIV time course data was examined for other genes induced with the same kinetics as the nine identified above, both by eye and by using the 'find similar' function of GeneSpring. Such kinetics immediately identified a four-gene operon induced very strongly in MIV (600-fold, HI0938-41). Examination of sequences upstream of its promoter revealed a CRE sequence that had been missed by previous searches because one base was specified only as 'K' (G or T); resequencing showed this to be an A, giving a perfect match to the CRE core consensus.

In parallel, the nine confirmed CRE sequences (excluding HI0938) were used to refine the CRE consensus, and the program RSA-tools (27) was then used to search the non-coding genome sequences for additional elements fitting it. A search using a stringent cutoff score of 20 returned eight of the nine input CREs, three additional CREs, and six sequences matching the CRP consensus. (The *comA* CRE was missed by this search because it overlaps an upstream coding region.) Examination of array data showed that the three genes with previously unrecognized CREs (*HI0659/0660*, *HI0952* and *HI1631*) were induced by MIV-starvation in the same manner as the confirmed CRE transcription units. Reducing the RSA-tools stringency cutoff to 10 and extending the search into upstream coding sequences produced many more CRP sites but only two more genes with candidate CREs (*mobB* and *gyrB*). Both these genes lack competence-specific regulation (induction in MIV or dependence on cAMP or Sxy; see below), and neither CRE is expected to have a strong influence on its gene's transcription–the *mobB* CRE because it is more than 300 bp upstream of the transcription start site, and the *gyrB* CRE because it is a poor match to the consensus. Expression levels of genes transcribed convergently with the CRE-regulon genes were also checked, to ensure that strong signals were not created by antisense transcription extending from convergently transcribed genes.

Properties of the complete CRE regulon are summarized in Fig. 2.3. The green bars on the right show how strongly each gene is induced, with the length of each bar indicating the ratio of maximum expression in MIV to t=0 expression in sBHI (for multi-gene operons the ratio shown is for the first gene).

On the left of Fig. 2.3 are the 13 CRE sequences and the distance of each from its likely promoter. The consensus of the 13 CRE sequences is shorter than but otherwise identical to that originally proposed (26). The sequence logos shown in Figure 2.4 reveal that CRE sites differ most strongly from the consensus of *H. influenzae* CRP sites in having G and C rather than T and A at the highly conserved symmetric positions 6 and 17; these are the positions where CRP bends DNA (28). CRE sequences are also less variable than CRP sequences, especially in the strings of Ts and As that flank the core. Consistent with the hypothesis that CRP binds to CRE sites, promoter-CRE spacing obeys the same constraints as promoter-CRP spacing in *E. coli* (29). The sequences upstream of each of the CRE-regulon genes were examined for additional motifs, using the programs Consensus (30), Gibbs recursive sampler (31) and Bioprospector (32). No new patterns were found, suggesting that CRE sites are likely to be the only sites where competence-regulatory factors bind DNA.

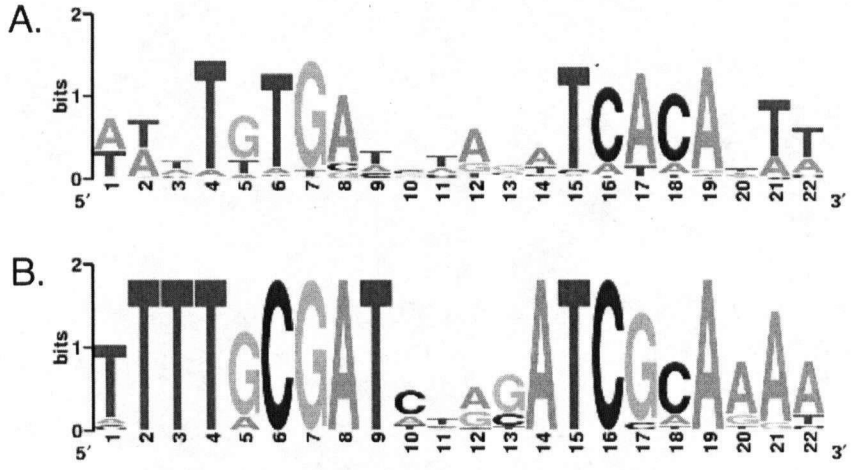


Figure 2.4 Sequence logo comparison of *H. influenzae* CRP and CRE consensus sequences.
CRP logo generated from the 45 CRP sites regulating the MIV-induced genes that are regulated by CRP but not Sxy. **B.** CRE logo from the 13 CRE sites regulating the genes of the CRE regulon.

CRP and cAMP regulate expression of the CRE regulon

CRP is the best candidate for the factor that binds CRE sites, as it and cAMP are known to regulate transcription of *comA* and *dprA* and to be essential for competence. Do CRP and cAMP also regulate transcription of all the other CRE-regulon genes? This was examined using arrays where RNA from MIV-induced *cya* cells was competed with RNA from a parallel culture incubated in cAMP-supplemented MIV. The *cya* mutant does not develop competence in

unsupplemented MIV but becomes fully competent if cAMP is provided (20). The yellow bars in Figure 2.3 show the degree to which expression of the first gene in each transcription unit depended on cAMP, calculated as the ratio of expression with cAMP to expression without cAMP. The cAMP dependence of all the CRE genes' expression was roughly proportional to their levels of induction in MIV (green bars and data not shown).

To test whether a CRE is in fact a CRP-binding site, we purified native CRP from *E. coli* and used electrophoretic mobility-shift assays to measure DNA-binding specificity. Bandshifts were apparent with both the *mglBAC* (CRP-binding site) and *comA-F* (CRE) promoter regions in reactions containing 5 to 500nM CRP and were detectable with 0.5nM CRP (Fig. 2.5). A DNA fragment without a CRP-binding site showed no bandshift even with 500nM CRP. The relative affinity of CRP for different bait DNAs was estimated by adding increasing concentrations of CRP into binding reactions. Comparison of lanes in which about half of the bait DNA was shifted (Fig. 2.5, 50 nM CRP for *comA-F*; 5 nM CRP for *mglBAC*;), revealed that CRP bound the *mglBAC* promoter with around 10-fold greater affinity than it bound the *comA-F* promoter. This finding is consistent with the 80-fold greater affinity of CRP for a synthetic (perfect) CRP-binding site over the same sequence with CRE-like G:C substitutions at base pairs 6 and 17 (33).

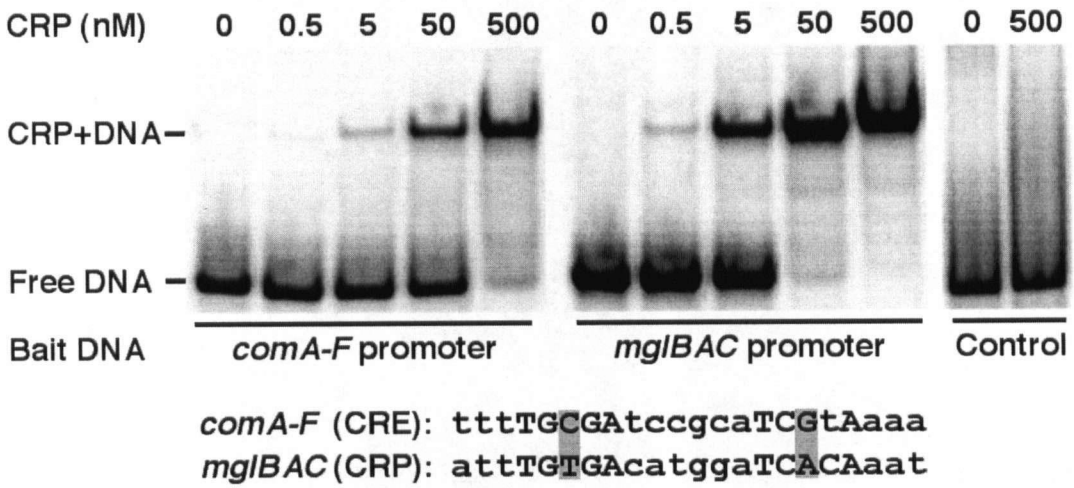


Figure 2.5 Electrophoretic mobility shift analysis of CRP-DNA complexes.
Bait DNAs are 130bp fragments amplified from *H. influenzae* chromosomal DNA; control DNA lacks an apparent CRP-binding site. CRP-binding site alignment: capital letters indicate agreement with the highly conserved regions of CRP-binding sites, grey boxes highlight the distinguishing bases of CRE sites

Sxy regulates expression of the CRE regulon

A *sxy* knockout mutant is unable to become competent and fails to induce expression of *comF* and *dprA lacZ*-fusions (5, 23-25). To find out whether Sxy controls all CRE regulon genes, and whether it controls other genes, we used microarrays to compare MIV-induced gene expression in cells carrying a *sxy* knockout (23) with that in wildtype cells. Except for *ssb* (see below), genes in the CRE regulon were expressed at 20-200-fold lower levels in the mutant. The purple bars in Fig. 2.3 show the magnitude of the Sxy dependence for the first gene in each transcription unit. Although the level of *comF* transcripts was too low to reliably measure, Zulty *et al.* have previously shown that Sxy is needed for *comF* expression (24); we have now confirmed this with quantitative PCR. With one exception, expression of genes lacking CRE sites was not changed by deletion of *sxy*, indicating that Sxy's only role may be to regulate the CRE regulon genes

The one exception was the operon containing genes HI0658-0654. These are moderately competence-induced (4-10-fold) with the same kinetics as the CRE regulon genes, and this induction depends on cAMP and Sxy. However, no CRE sequence could be identified in the 160bp noncoding region upstream of HI0658, and none of the genes' functions are obviously related to DNA uptake. This operon is directly downstream of the CRE-regulated HI0660-0659 operon (Fig. 2.3), so transcription may read through from it into HI0658-0654 (the 160 intervening bp lack any obvious transcriptional terminator). Comparisons of HI0658 and HI0659 to homologs in *Actinobacillus pleuropneumoniae* and *Mannheimia haemolytica* confirmed the stop and start codon assignments.

Other starvation-induced genes

To ensure that no other competence genes had been overlooked, all genes at least four-fold induced in MIV were examined for function and for dependence on cAMP and Sxy (Fig. 2.6). Most of these 151 genes were found to be CRP dependent, consistent with the evidence that cAMP levels rise during competence induction and with the large number of CRP-regulated genes postulated by Tan *et al.* (21). The induced genes included 23 of the 25 genes in the CRE regulon, (*ssb* and *comF* did not meet the 4-fold induction criterion), two genes in the PurR regulon, one in the TrpR regulon, and 81 other genes with CRP sites and cAMP-dependent, *sxy*-independent expression in MIV (CRP-regulon genes). The other CRP-regulon genes fell into several groups: 27 genes of unknown functions, 23 genes involved in sugar utilization, and 31

other genes mainly with roles in nutrient uptake and central metabolism. None of these non-CRE genes has been implicated in DNA uptake.

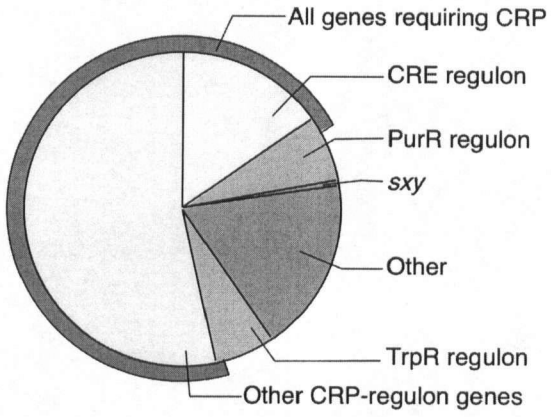


Figure 2.6 Starvation-induced genes.
Categorization of the 151 genes induced at least 4-fold on transfer to MIV.

Forty four of the MIV-induced genes depended on neither CRP nor Sxy. In addition to *sxy* itself (discussed below), these included the PurR-regulon and TrpR-regulon genes shown in Fig. 2.2C. Although genes for synthesis of other amino acids were not induced, transfer to MIV caused rapid induction of genes for tryptophan biosynthesis, presumably due to the lack of tryptophan in the casamino acid component of MIV. Supplementation of MIV with tryptophan did not affect competence development (data not shown). Genes in the purine biosynthetic pathway were also rapidly induced, confirming that transfer to MIV causes rapid depletion of purine pools. Genes for pyrimidine synthesis (*pyrD, E, F* and *G*) were expressed quite strongly during exponential growth in sBHI and were not further induced by transfer to MIV (Fig. 2.2C). The final categories of MIV-induced genes comprise 15 genes whose functions have no obvious connection to competence, and 13 genes whose functions are unknown. The lack of any competence-related genes in this set suggests that the CRE regulon includes all of the genes that need to be induced for competence development.

Genes down-regulated in MIV

Many of the 44 genes down-regulated at least four-fold play roles in translation. Transcription of the 29 genes in the two ribosomal protein operons (HI0776-HI-786 and HI0788-HI0803) was reduced transiently, though not all genes met the four-fold-reduction cutoff. The conserved operon containing *nusA* and *infB* was also down regulated, as was *rpoBC*. The other down-

regulated genes did not fall into any evident groups; *cya* was the only downregulated gene with a known connection to competence.

Expression of regulatory genes

Regulation of CRP and cAMP.

Unlike its *E. coli* homolog, the *H. influenzae crp* promoter has no CRP sites; consistent with this, microarrays showed that *crp* mRNA was only weakly induced by transfer of cells to MIV (Fig. 2.2D) and unaffected by mutation of *cya*. The *H. influenzae cya* gene, like its *E. coli* homolog, has a good CRP site overlapping its promoter, which is predicted to act as a repressor rather than an activator of transcription (20). Consistent with this, *cya* mRNA was sharply decreased after transfer to MIV (Fig. 2.2D) and increased about 6-fold in *cya* mutant cells. The *icc* gene (cAMP phosphodiesterase) was induced 5-6-fold in MIV (Fig. 2.2D) and decreased about 2-fold in the absence of cAMP. The induction of *icc* in MIV would increase cAMP turnover and, with decreased transcription of *cya*, limit the cell's long-term response to activation of adenylate cyclase by the PTS. The *sxy* knockout mutation had no effect on transcription of *cya*, *icc* or *crp*.

Regulation of Sxy.

Expression of the regulatory gene *sxy* was induced 16-40-fold after transfer to MIV, with maximum expression in the 30 minute sample (Fig. 2.2D). This induction is consistent with Sxy's role as a positive regulator of CRE-regulon genes, and with previous primer-extension analysis (24). The *cya* mutation had no effect on *sxy* expression, contrary to a previous report that cAMP induces *sxy* transcription (24, 25). Expression of a *lacZ* fusion to the *sxy* promoter does not depend on the presence of an intact *sxy* gene, so transcriptional autoregulation is not a factor (25).

Competence development in rich medium

Cultures become modestly competent in colonies on sBHI agar plates and when liquid sBHI cultures approach stationary phase (23). With the exception of *ssb*, all genes in the CRE regulon were also modestly induced (4-20x) as stationary phase approached during growth in rich medium (data not shown). This suggests that the low level of competence seen at this stage is not due to failure of a particular competence function, but to a general low induction of all

components. This is supported by the modest increases in expression of *sxy* and of the CRP-regulon genes induced in MIV.

DISCUSSION

What do the CRE-regulon genes do?

DNA uptake and translocation functions:

Several of the genes in the CRE regulon are known to have roles in assembly of the uptake machinery or in DNA transport. Insertions disrupting *comA* and *comC* prevent DNA binding and uptake; however their mutant phenotypes could be due to polar effects on *comE*. ComA is predicted to be cytoplasmic and ComC to be targeted to the inner membrane. ComE is a member of the secretin family of gated pore proteins associated with Type IV pili. The *pilA* gene encodes a typical pilin subunit of Type IV pili; the *pilBCD* genes are homologous to genes for pilin processing and pilus assembly. As *H. influenzae* lacks visible Type IV pili these genes likely produce a short pseudopilus. An insertion disrupting the inner-membrane protein Rec-2 allows DNA binding and uptake into the periplasm but the DNA cannot be translocated into the cytoplasm (3). Mutations in *comF* (original name *com101A*) cause a similar phenotype (38). The ComE1 protein is homologous to the C-terminal region of the *Bacillus subtilis* DNA-uptake protein ComEA, and our preliminary data implicates ComE1 in DNA uptake by *H. influenzae* (S. Molnar and R. Redfield, manuscript in preparation).

Genes in the HI0938-0941 operon have not previously been associated with competence. They have good homologs only in the Pasteurellaceae, but weak homologs occur in similar operons in many other bacteria. All are predicted to be secreted from the cytoplasm, and an insertion in HI0938 prevents DNA uptake. (S. Molnar and R. Redfield, manuscript in preparation). HI1182/1183 (incorrectly annotated as two ORFs due to a sequencing frameshift) belongs to a small family of ATP-dependent DNA ligase with signal sequences for secretion into the periplasm. Both the *H. influenzae* and *N. gonorrhoeae* ligases have been well-characterized *in vitro*, but no periplasmic function is known (39, 40).

Cytoplasmic functions:

The CRE-regulon proteins with known cytoplasmic functions all interact with DNA but have not been implicated in DNA uptake. Insertions in *comM* and *dprA* cause DNA entering the cytoplasm to be degraded before it can recombine with the chromosome (4, 6). ComM is a member of the YifB subfamily of AAA-ATPase proteins - its possession of a *lon* protease

domain suggests it may be an ATP-dependent protease (41). DprA (Smf) is also predicted to bind ATP; it is required for transformation in a number of bacteria and its homolog in *S. pneumoniae* also protects DNA from degradation (42). SSB homologues are ubiquitous and well characterized. Its ability to bind and stabilize single-stranded DNA is essential for DNA replication and repair, and also plays a role in homologous recombination (43). Culturing cells in MIV had only modest effects on *ssb* expression, perhaps because *ssb* transcripts are abundant in log-phase cells. The function of *radC* is less well understood; it encodes a RecG-like protein thought to function at stalled DNA replication forks, and is also a component of the *S. pneumoniae* and *B. subtilis* competence regulons (13, 44, 45).

Other proteins:

Other CRE regulon genes encode cytoplasmic proteins of unknown function. *HI0365* contains a Fe-S oxidoreductase domain. *HI0659* and *HI0660* are short proteins that are conserved as an operon in a number of distantly-related bacteria; HI0659 contains a helix-turn-helix domain in the XRE family (46). *HI1631* has no known homologs.

In *H. influenzae* the ability to take up DNA develops as a unified response to changing conditions. All competence genes induced in MIV starvation medium are regulated by Sxy, suggesting that the regulation of competence will be understood only when we understand the regulation of *sxy* expression and its role in expression of the CRE-regulon genes.

This work identified 11 new genes not previously associated with competence. The most intriguing of these is the periplasmic ATP-dependent DNA ligase encoded by HI1182/83. Similar secreted ATP-dependent DNA ligases are found in *Neisseria* and several other bacteria (39, 40); they belong to one of two newly-discovered families of bacterial ATP-dependent DNA ligases. Both the *H. influenzae* and *Neisseria* ligases are known to seal nicks but not blunt ends, and Magnet and Blanchard presciently speculated that they might function in competence (40). Consistent with this, transformation with cloned fragments bearing restriction fragment ('sticky') ends often gives transformants containing conjoined DNA fragments, and a periplasmic ligase activity was proposed in explanation (48). However, there is no obvious role for ligation in DNA uptake, and there is unlikely to be any ATP in the periplasm, especially because one of the MIV-induced genes is the periplasmic 5'-nucleotidase encoded by HI0206.

We do not know whether the remaining ten genes of the CRE regulon all contribute to competence or reflect a broader role of the CRE regulon, perhaps resolving other problems

created by depletion of nucleotide pools. In *Bacillus subtilis* and *Streptococcus pneumonia*, what were originally thought to be competence-specific signals control many genes of diverse function (12, 13, 45). Competence regulons have not yet been identified in other Gram-negative bacteria. Characterization of the functions of the new CRE-regulon genes and of possible CRE regulons in related bacteria should help resolve this issue.

REFERENCES

1. 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.
2. Larson,T.G. and Goodgal,S.H. (1991) Sequence and transcriptional regulation of com101A, a locus required for genetic transformation in Haemophilus influenzae. *J. Bacteriol.,* **173**, 4683-4691.
3. McCarthy,D. (1989) Cloning of the rec-2 locus of Haemophilus influenzae. *Gene,* **75**, 135-143.
4. 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.
5. Karudapuram,S. and Barcak,G.J. (1997) The Haemophilus influenzae dprABC genes constitute a competence-inducible operon that requires the product of the tfoX (sxy) gene for transcriptional activation. *J. Bacteriol.,* **179**, 4815-4820.
6. 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.
7. Tomb,J.F., Barcak,G.J., Chandler,M.S., Redfield,R.J. and Smith,H.O. (1989) Transposon mutagenesis, characterization, and cloning of transformation genes of Haemophilus influenzae Rd. *J. Bacteriol.,* **171**, 3796-3802.
8. Dougherty,B.A. and Smith,H.O. (1999) Identification of Haemophilus influenzae Rd transformation genes using cassette mutagenesis. *Microbiology,* **145**, 401-409.
9. Gwinn,M.L., Stellwagen,A.E., Craig,N.L., Tomb,J.F. and Smith,H.O. (1997) In vitro Tn7 mutagenesis of Haemophilus influenzae Rd and characterization of the role of atpA in transformation. *J. Bacteriol.,* **179**, 7315-7320.
10. Beattie,K.L. and Setlow,J.K. (1971) Transformation-defective strains of Haemophilus influenzae. *Nat New Biol,* **231**, 177-179.
11. Caster,J.H., Postel,E.H. and Goodgal,S.H. (1970) Competence mutants: isolation of transformation deficient strains of Haemophilus influenzae. *Nature,* **227**, 515-517.
12. Dagkessamanskaia,A., Moscoso,M., Henard,V., Guiral,S., Overweg,K., Reuter,M., Martin,B., Wells,J. and Claverys,J.P. (2004) Interconnection of competence, stress and CiaR regulons in Streptococcus pneumoniae: competence triggers stationary phase autolysis of ciaR mutant cells. *Mol. Microbiol.,* **51**, 1071-1086.
13. Peterson,S.N., Sung,C.K., Cline,R., Desai,B.V., Snesrud,E.C., Luo,P., Walling,J., Li,H., Mintz,M., Tsegaye,G., Burr,P.C., Do,Y., Ahn,S., Gilbert,J., Fleischmann,R.D. and Morrison,D.A. (2004) Identification of competence pheromone responsive genes in Streptococcus pneumoniae by use of DNA microarrays. *Mol. Microbiol.,* **51**, 1051-1070.
14. Ogura,M., Yamaguchi,H., Kobayashi,K., Ogasawara,N., Fujita,Y. and Tanaka,T. (2002) Whole-genome analysis of genes regulated by the Bacillus subtilis competence transcription factor ComK. *J. Bacteriol.,* **184**, 2344-2351.

15. Reizer,J. and Reizer,A. (1996) A voyage along the bases: novel phosphotransferase genes revealed by in silico analyses of the Escherichia coli genome. *Res Microbiol,* **147**, 458-471.

16. Macfadyen,L.P., Dorocicz,I.R., Reizer,J., Saier,M.H.J. and Redfield,R.J. (1996) Regulation of competence development and sugar utilization in Haemophilus influenzae Rd by a phosphoenolpyruvate:fructose phosphotransferase system. *Mol. Microbiol.,* **21**, 941-952.

17. Macfadyen,L.P. (1999) PTS regulation of competence in Haemophilus influenzae. PhD Thesis, University of British Columbia.

18. Gwinn,M.L., Yi,D., Smith,H.O. and Tomb,J.F. (1996) Role of the two-component signal transduction and the phosphoenolpyruvate: carbohydrate phosphotransferase systems in competence development of Haemophilus influenzae Rd. *J. Bacteriol.,* **178**, 6366-6368.

19. Chandler,M.S. (1992) The gene encoding cAMP receptor protein is required for competence development in Haemophilus influenzae Rd. *Proc Natl Acad Sci U S A,* **89**, 1626-1630.

20. Dorocicz,I.R., Williams,P.M. and Redfield,R.J. (1993) The Haemophilus influenzae adenylate cyclase gene: cloning, sequence, and essential role in competence. *J. Bacteriol.,* **175**, 7142-7149.

21. 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.

22. Wise,E.M.J., Alexander,S.P. and Powers,M. (1973) Adenosine 3':5'-cyclic monophosphate as a regulator of bacterial transformation. *Proc Natl Acad Sci U S A,* **70**, 471-474.

23. 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.

24. Zulty,J.J. and Barcak,G.J. (1995) Identification of a DNA transformation gene required for com101A+ expression and supertransformer phenotype in Haemophilus influenzae. *Proc Natl Acad Sci U S A,* **92**, 3616-3620.

25. Bannister,L.A. (1999) An RNA secondary structure regulates sxy expession and competence development in Haemophilus influenzae. PhD Thesis, University of British Columbia.

26. Macfadyen,L.P. (2000) Regulation of competence development in Haemophilus influenzae. *J Theor Biol,* **207**, 349-359.

27. van Helden,J. (2003) Regulatory sequence analysis tools. *Nucleic Acids Res.,* **31**, 3593-3596.

28. 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.

29. Kolb,A., Busby,S., Buc,H., Garges,S. and Adhya,S. (1993) Transcriptional regulation by cAMP and its receptor protein. *Annu. Rev. Biochem.,* **62**, 749-795.

30. Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics,* **15**, 563-577.

31. Thompson,W., Rouchka,E.C. and Lawrence,C.E. (2003) Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Res.,* **31**, 3580-3585.

32. 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.

33. 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.

34. Spiro,S., Gaston,K.L., Bell,A.I., Roberts,R.E., Busby,S.J. and Guest,J.R. (1990) Interconversion of the DNA-binding specificities of two related transcription regulators, CRP and FNR. *Mol. Microbiol.,* **4**, 1831-1838.

35. Chandler,M.S. and Smith,R.A. (1996) Characterization of the Haemophilus influenzae topA locus: DNA topoisomerase I is required for genetic competence. *Gene,* **169**, 25-31.

36. Finkel,S.E. and Kolter,R. (2001) DNA as a nutrient: novel role for bacterial competence gene homologs. *J. Bacteriol.,* **183**, 6288-6293.

37. Ma,C. and Redfield,R.J. (2000) Point mutations in a peptidoglycan biosynthesis gene cause competence induction in Haemophilus influenzae. *J. Bacteriol.,* **182**, 3323-3330.

38. Larson,T.G. and Goodgal,S.H. (1992) Donor DNA processing is blocked by a mutation in the com101A locus of Haemophilus influenzae. *J. Bacteriol.,* **174**, 3392-3394.

39. Cheng,C. and Shuman,S. (1997) Characterization of an ATP-dependent DNA ligase encoded by Haemophilus influenzae. *Nucleic Acids Res.,* **25**, 1369-1374.

40. Magnet,S. and Blanchard,J.S. (2004) Mechanistic and kinetic study of the ATP-dependent DNA ligase of Neisseria meningitidis. *Biochemistry,* **43**, 710-717.

41. Iyer,L.M., Leipe,D.D., Koonin,E.V. and Aravind,L. (2004) Evolutionary history and higher order classification of AAA+ ATPases. *J. Struct. Biol.,* **146**, 11-31.

42. 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.

43. Raghunathan,S., Kozlov,A.G., Lohman,T.M. and Waksman,G. (2000) Structure of the DNA binding domain of E. coli SSB bound to ssDNA. *Nat. Struct. Biol.,* **7**, 648-652.

44. 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.

45. Berka,R.M., Hahn,J., Albano,M., Draskovic,I., Persuh,M., Cui,X., Sloma,A., Widner,W. and Dubnau,D. (2002) Microarray analysis of the Bacillus subtilis K-state: genome-wide expression changes dependent on ComK. *Mol. Microbiol.,* **43**, 1331-1345.

46. Wood,H.E., Devine,K.M. and McConnell,D.J. (1990) Characterization of a repressor gene (xre) and a temperature-sensitive allele from Bacillus subtilis prophage, PBSX. *Gene,* **96**, 83-88.

47. Nudler,E. and Mironov,A.S. (2004) The riboswitch control of bacterial metabolism. *Trends Biochem. Sci.,* **29**, 11-17.

48. Poje,G. and Redfield,R.J. (2003) Transformation of Haemophilus influenzae. *Methods Mol Med,* **71**, 57-70.

49. Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.F., Dougherty,B.A., Merrick,J.M. and et,a. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science,* **269**, 496-512.

50. Rozen,S. and Skaletsky,H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.,* **132**, 365-386.

51. DeRisi,J.L., Iyer,V.R. and Brown,P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science,* **278**, 680-686.

52. Chen,W.J., Gross,L., Joho,K.E. and McAllister,W.T. (1992) A modified kanamycin-resistance cassette to facilitate two-codon insertion mutagenesis. *Gene,* **111**, 143-144.

53. Poje,G. and Redfield,R.J. (2003) General methods for culturing Haemophilus influenzae. *Methods Mol Med,* **71**, 51-56.

54. Bacon,J., James,B.W., Wernisch,L., Williams,A., Morley,K.A., Hatch,G.J. and *al.,e.* (2004) The influence of reduced oxygen availability on pathogenicity and gene expression in Mycobacterium tuberculosis. *Tuberculosis (Edinb),* **84**, 205-217.

55. Zhang,X.P., Gunasekera,A., Ebright,Y.W. and Ebright,R.H. (1991) Derivatives of CAP having no solvent-accessible cysteine residues, or having a unique solvent-accessible cysteine residue at amino acid 2 of the helix-turn-helix motif. *J Biomol Struct Dyn,* **9**, 463-473.

56. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.,* **14**, 1188-1190.

CHAPTER THREE

Non-canonical CRP sites control competence regulons in *Escherichia coli* and many other Gamma-proteobacteria.[2]

INTRODUCTION

The *E. 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 22bp DNA sites with the consensus half site 5'-$A_1A_2A_3$**$T_4G_5T_6G_7A_8$**$T_9C_{10}T_{11}$. 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).

---

[2] A version of this chapter has been published. Cameron A.D.S. and Redfield R.J. (2006) Non-canonical CRP sites control competence regulons in Escherichia coli and many other γ-proteobacteria. *Nucleic Acids Res.* **34:** 6001-6014

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 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 22nt consensus and all but 9 sites are mismatched at one or more positions of the 10nt 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'-$T_1T_2T_3\mathbf{T_4G_5C_6G_7A_8}T_9C_{10}T_{11}$ (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 (*V. 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 (for example (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, *Haemophilus influenzae KW20 Rd*, *Haemophilus ducreyi 35000HP*, *Mannheimia succiniciproducens MBEL55E*, *Pasteurella multocida Pm70*, *Pseudomonas aeruginosa PAO1*, *Pseudomonas fluorescens Pf-5*, *Salmonella typhimurium LT2 SGSC1412*, *Vibrio 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 (*P. syringae*, *P. fluorescens PfO-1*, *P. putida KT2440*, *P. syringae phaseolicola 1448A*, and *P. syringae pv B728a*) and five additional *Xanthomonadaceae* genomes (*X. citri*, *X. campestris 8004*, *X. campestris vesicatoria 85-10*, *X. fastidiosa Temecula1*, *X. 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 analyzed all paralog promoters. For the *Pseudomonadaceae* and *Xanthomonadaceae* species not listed in Figure 3.1, 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 –300bp 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 analyzed 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://bioprospector.stanford.edu/cgi-bin/BPsearch.pl. 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 3.1. 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 previously described (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 LB (30μg/ml chloramphenicol or 10μg/ml kanamycin). Knockout strains were made chemically competent with RbCl and transformed with pASKA*sxy* as previously described (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 (25μg/ml kanamycin and 100μg/ml ampicillin) and *crp* expression was induced at OD_{600} 0.6 with 1mM IPTG. Cells were harvested after 4.5hr by centrifugation and the pellet were frozen overnight at -20°. Native CRP was purified as follows: the pellet was resuspended in lysis buffer (50mM sodium phosphate, 300mM sodium chloride, 10mM imidazole), then treated with 1mg/ml lysozyme for 30min at 24° followed by sonication on ice. Insoluble material was removed by centrifugation at 10,000g for 25min and the supernatant was

then incubated with nickel-nitriloacetic acid agarose beads for 1hr at 4° with gentle rocking. The agarose beads were loaded in a column and washed twice with four column volumes of wash buffer (50mM sodium phosphate, 300mM sodium chloride, 20mM imidazole), and protein was collected in elution buffer (50mM sodium phosphate, 300mM sodium chloride, 250mM imidazole). Purified protein was desalted with Nanosep 3K Omega membranes (Pall), then resuspended in storage buffer (20% glycerol, 40mM Tris, 200mM 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'-CGCTGTAAATCTGCATCGGA, yrfDR 5'-CAGTCTGTTGCATTCTGCTGGG, lacZF 5'-GCACGACAGGTTTCCCGACT, lacZR 5'-CACAATTCCACACAACATAC, hofBF 5'-GCCTACCGCATCCGCTT, hofBR 5'-CCAGGTTTCCCAGCACTTTTAAT. Amplicons were purified using polyacrylamide gel electrophoresis. Bands were then excised and DNA was eluted from macerated gel overnight in TE at 37°, ethanol precipitated and resuspended in 10mM Tris. DNA was end-labeled with T4 polynucleotide kinase using a ten fold molar excess of γ-^{32}P ATP, and unincorporated label was removed with a PCR cleanup spin column (Sigma).

CRP-DNA binding reactions (10µl) contained 100nM CRP, 10mM Tris (pH 8.0), 50mM KCl, 5% (v/v) glycerol, 250 µg/ml bovine serum albumin, 100µM cAMP, 1 mM dithiothreitol, 40µg/µl poly(dI-dC) DNA, and 1x10^{6}cpm labeled bait DNA. Reactions were incubated at room temperature for ten minutes before being loaded onto a prerun polyacrylamide gel (30:1 acrylamide/bisacrylamide; 0.2xTBE (89 mM Tris, 89 mM borate, 2 mM EDTA (pH 8.3)), 2% glycerol, and 200 µM cAMP; running buffer 0.2xTBE and 200 µM cAMP. After electrophoresis for 2.5 hours at 100V, the gel was dried and exposed for two hours 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 (1xTAE). 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'-GCATACTGACCACCTGTAATTTCTG, mglBF 5'-GTCCAGCATTCCGGTGTTTGG, 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 1/1,000 and 1/10,000; 23SF 5'-GCTGATACCGCCCAAGAGTT, 23SR 5'-CAGGATGTGATGAGCCGAC. 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 3.8 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 homologs 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 analyzed here have well

resolved phylogenies (see tree on the left side of Figure 3.1) and are used routinely to represent the diversity of γ-proteobacteria (33-35).

Figure 3.1 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.

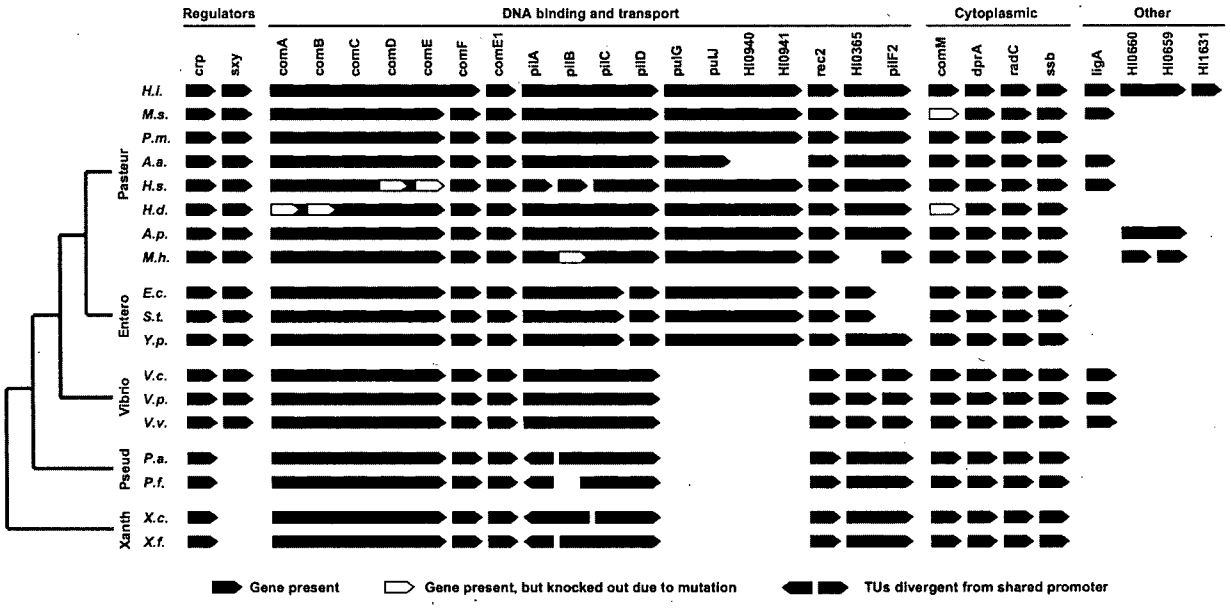


Figure 3.1 Orthologs of *H. influenzae* CRP-S-regulated genes in other γ-proteobacteria. Solid lines depict transcriptional units (gene lengths not to scale). Cladogram adapted from Lerat *et al.* (33). Abbreviations: Pasteur, *Pasteurellaceae*; Entero, *Enterobacteriaceae*; Vibrio, *Vibrionaceae*; Pseud, *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*.

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 3.1 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 3.1 (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 Appendix 2; below we present only sequence logo versions of the shared motifs.

Table 3.1 Details of phylogenetic footprinting.

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

* Includes only *H. influenzae*, *M. succiniciproducens*, *P. multocida* and *H. ducreyi* promoters.
Grey background highlights the results of an alternate search strategy employed for *Vibrionaceae* (explained in Results)

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

Phylogenetic footprint analysis of the 91 *Pasteurellaceae* promoters in Figure 3.1 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 Fig. 3.2A; the sites themselves are listed in Appendix 2, 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 3.2A 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 Fig. 3.2B; sites listed in Appendix 2, Table 2). As expected, this motif strongly resembled CRP-N sites.

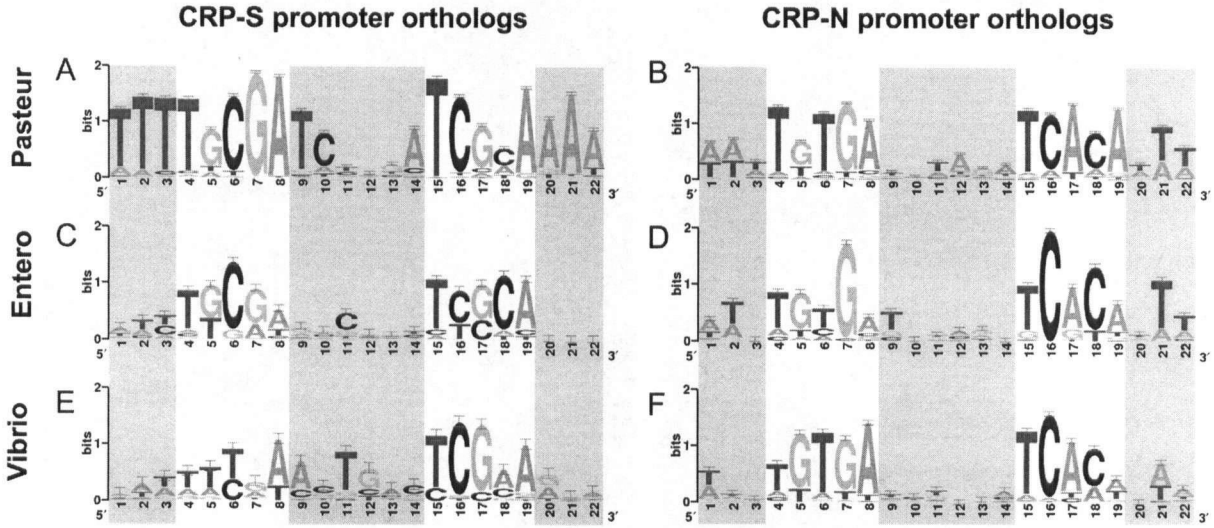


Figure 3.2 Motifs from pooled gene promoters.
A+B. *Pasteurellaceae*; **C+D.** *Enterobacteriaceae*; **E+F.** *Vibrionaceae*. CRP-S promoter orthologs are those in Figure 3.1. Logos were generated from alignment of all sites in Appendix 2 Tables 1 through 6 using WebLogo (http://weblogo.cbr.nrc.ca/logo.cgi). 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).

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 combined). The weight scores (I_{seq}) for all sites overlapped the scores of the *H. influenzae* CRP sites used to construct the matrix (Fig. 3.3A). 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 22bp 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.

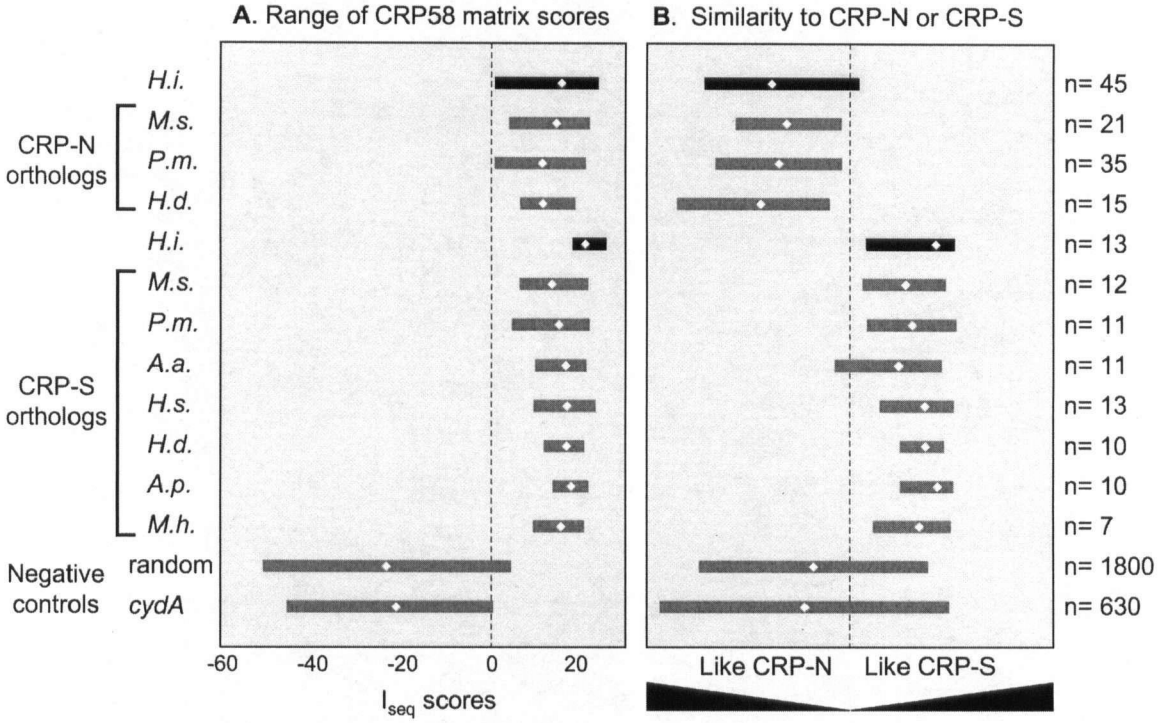


Figure 3.3 Similarity of putative CRP sites to experimentally determined sites.
Bars indicate range of scores (black bars are experimentally determined sites); white diamonds are the mean scores. **A.** Sites scored with CRP58 matrix. **B.** Scores indicate the difference of I_{seq} for each site scored with CRP45 (CRP-N) and CRE13 (CRP-S) matrices.

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 3.3B summarizes the scores of the *Pasteurellacean* promoters. All but one of the 74 predicted sites from *Pasteurellaceae* genes in Figure 3.1 (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.

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

Phylogenetic footprint analysis of the 33 *Enterobacteriaceae* promoters in Figure 3.1 (CRP-S orthologs) identified a single conserved motif present at 38 sites (summarized by the sequence logo in Fig. 3.2C; sites are listed in Appendix 2, Table 3). Analyzing the 90 promoters of orthologs of *H. influenzae* CRP-N-regulated genes yielded 57 sites sharing one motif (sequence logo in Fig. 3.2D; sites listed in Appendix 2, Table 4). As expected, the CRP-N-ortholog motif in Figure 3.2D is a canonical CRP site, whereas the CRP-S-ortholog promoter motif in Figure 3.2C has significant overrepresentation of the C_6 and G_{17} bases characteristic of CRP-S sites. Figure 3.4 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.

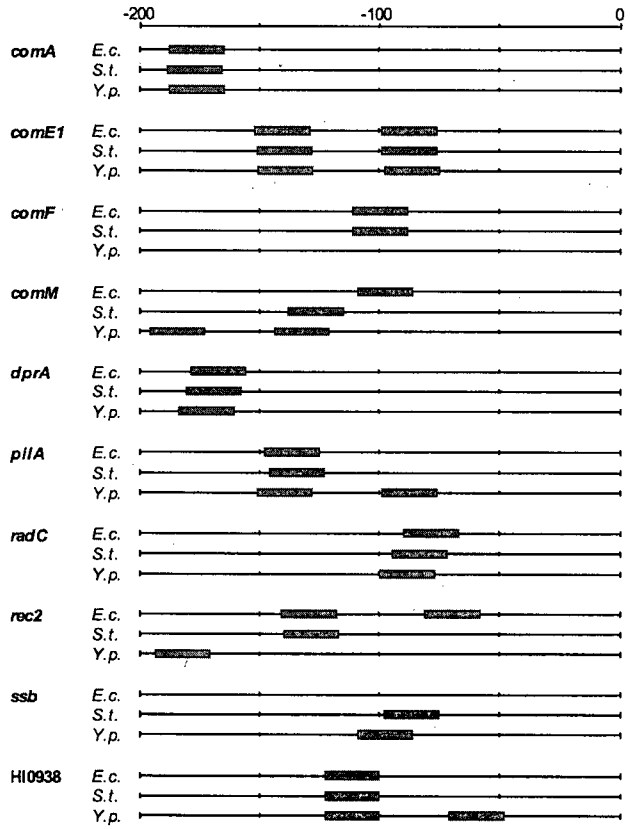


Figure 3.4 Physical map of *Enterobacteriaceae* promoters, named according to *H. influenzae* orthologs in Figure 3.1.
Grey boxes indicate positions of putative CRP-S sites relative to start codons (sites listed in Appendix 2, Table 3). In all three *comA* promoters, a second CRP-S lies >200bp away from the gene start (*E.c.* -246, *S.t.* -247, *Y.p.* -246).

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

Although *Vibrio 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 3.1 were analyzed 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 6 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 3.2E. This search identified a single motif present at 24 sites in the 15 promoters (sites listed in Appendix 2 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 3.2E, 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 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* (Fig. 3.2F), and one matching the PurR repressor binding site consensus (Fig. 3.5); the genes and sites are listed in Appendix 2 Tables 6 and 7. The CRP motif in Figure 3.2F 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*.

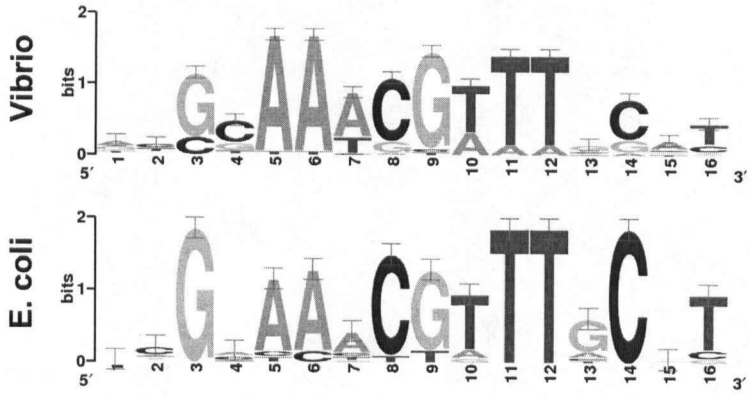


Figure 3.5 PurR binding site motifs.
PurR logos from alignment of 27 *Vibrionaceae* sites in Appendix 2, Table 7 and the 15 *E. coli* sites listed at RegulonDB.

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.c.*, 7 in *V.p.*, and 9 in *V.v.*), including 13 of those that also had CRP-N motifs (Appendix 2, 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 6 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 3.1 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 analyzed 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 to detect site-specific DNA binding (Fig. 3.6). 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 (not shown)

served as negative controls. No bandshifts were observed in the absence of CRP or with negative control *hofB* DNA (Fig. 3.6, 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 Fig. 3.6) — 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 3.6; the low affinity of CRP for the *ppdD* and *yrfD* CRP-S sites is consistent with their low scores.

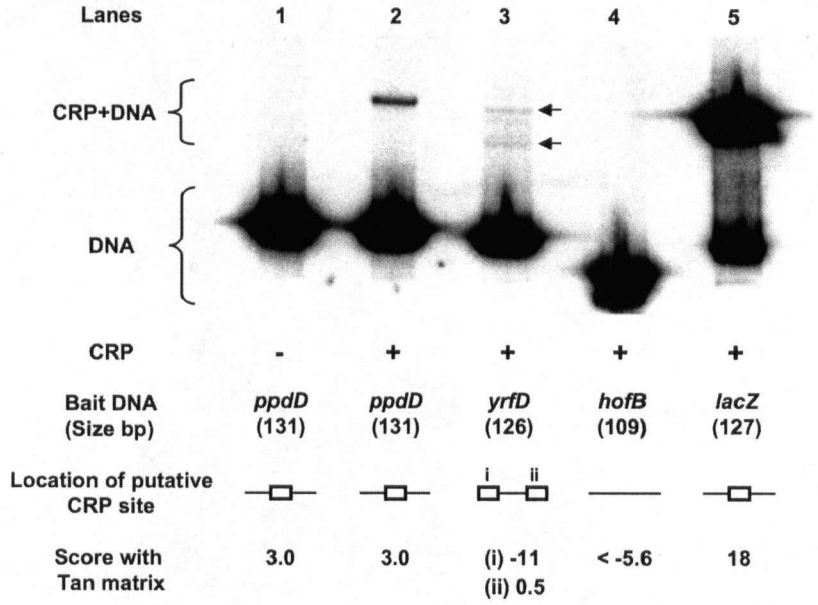


Figure 3.6 Electrophoretic mobility-shift assay.
E. coli CRP binds specifically to *E. coli* promoters containing putative CRP-S sites. Arrows indicate faint bandshifts with *yrfD* (b3395) promoter.

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)(Fig. 3.7A). 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 (Fig. 3.7B and C). The red bars in Figure 3.7B 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*.

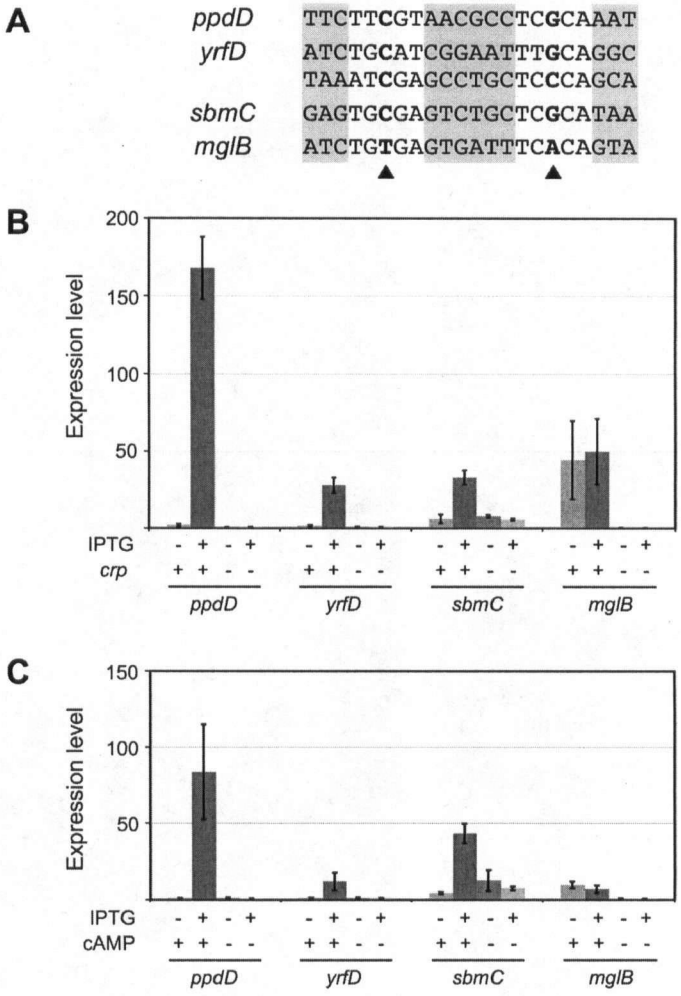


Figure 3.7 Sxy-dependent gene expression from *E. coli* CRP-S promoters measured using quantitative PCR.

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.

To test the CRP-dependence of these genes, transcription analysis was repeated using a host carrying a *crp* knockout (Fig. 3.7B). 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 to not 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 (Fig. 3.7C), 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 (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 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 EPV clade and its results disagreed with the established relationships presented in Figure 3.1. 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 3.8). Five lineages were resolved, and their congruence with the established bacterial phylogeny shown on the left of Figure 3.1 confirms the findings of Korner *et al.* (52) that CRP is ancestral to the γ-proteobacteria. The Sxy phylogeny in Figure 3.8 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, R185); Figure 3.8 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, 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.

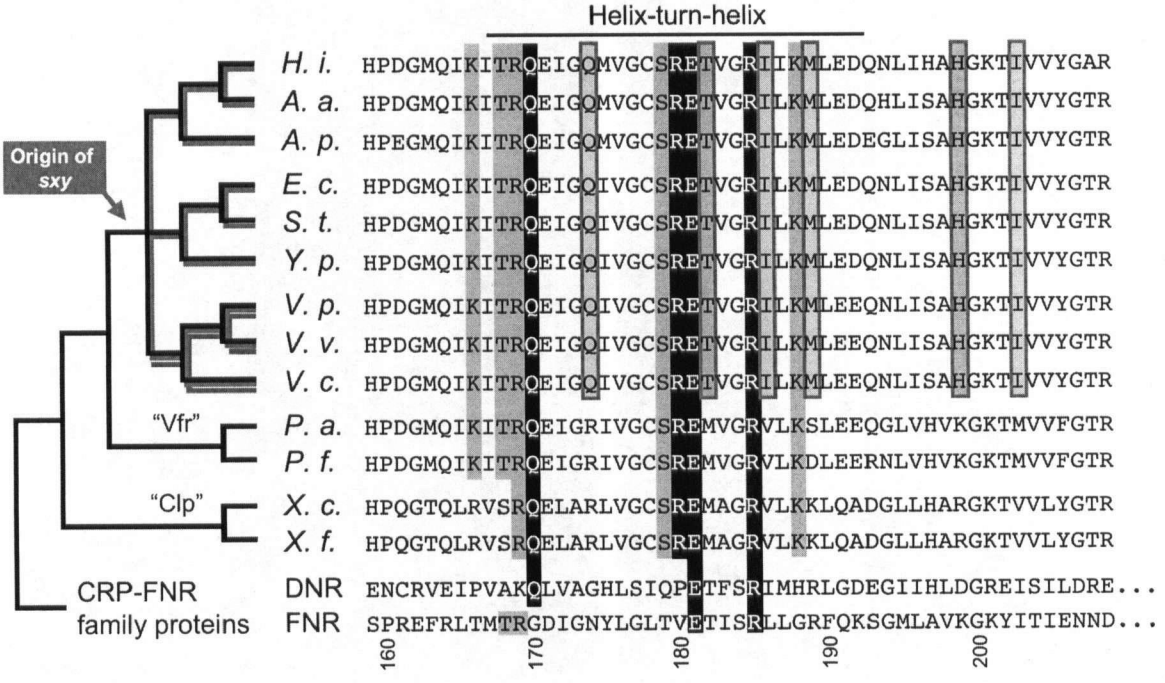


Figure 3.8 Evolutionary history of *crp* (black) and *sxy* (red).
Nodes of the phylogenetic tree are supported by bootstrap values over 80%, except for the root of the EPV clade where branching order could not be resolved due to low (<40%) bootstrap values. A gene duplication event generated *sxy* paralogs 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 3.1.

CRP's DNA-binding domain is contained within 50 C-terminal amino acids (aligned in Figure 3.8); 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 (Fig. 3.2A). 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 3.9 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). 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.

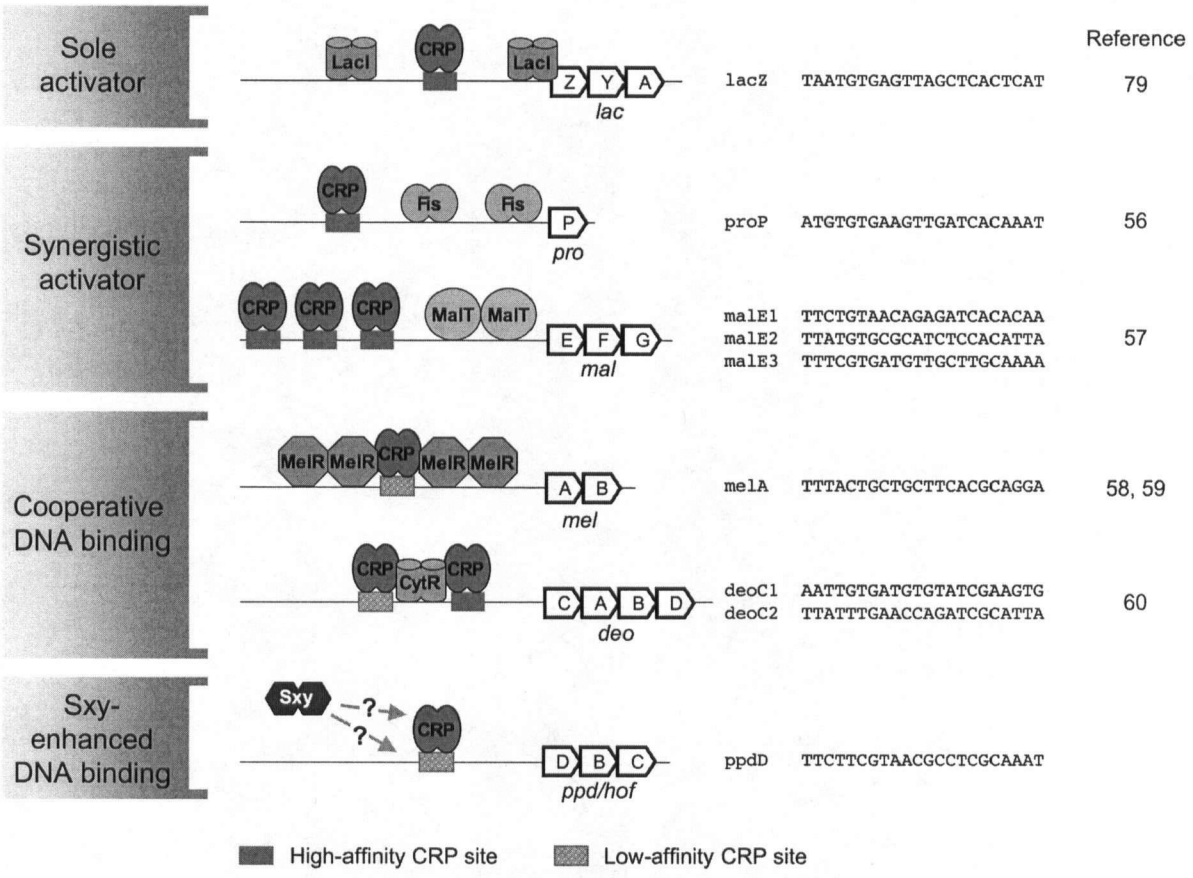


Figure 3.9 Categories of CRP-activated promoters in *E. coli*.
In promoters where CRP acts as a sole transcription activator or synergistic activator, it binds to high-affinity sites. CRP requires cooperativity with MelR and CytR to bind to low affinity sites in the *melAB* and *deoCABD* promoters, among others. Sxy is hypothesized to interact directly with CRP, and may also bind DNA.

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 (Appendix 2, 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 (for example, the *melA* and *deoC* sites in Figure 3.9).

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 (Fig. 3.4). *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 (Fig. 3.1). Moreover, in *H. somnus* the *pil* operon has dissociated into three 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 (Fig. 3.1). 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 (discussed in Chapter 5). 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 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 hundred 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 (Fig. 3.6) (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.

REFERENCES

1. Emmer,M., deCrombrugghe,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 U S A,* **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 Evolutionary Biology,* 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,* 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**, 2006.0008.

29. Ausubel,F.M. (1995) Current protocols in molecular biology. J. Wiley & Sons, Inc, Brooklyn, NY.

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, U.S.A.

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**, E19.

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 U S A,* **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 U S A,* **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 U S A,* **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 U S A,* **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 U S A*, **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 U S A*, **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**, D98-102.

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., et al. (ed.), *Escherichia coli and Salmonella typhimurium,* Washington, D.C., Vol. II, pp. 1232-1245.

CHAPTER FOUR

Sxy induces competence by enhancing CRP binding and transcription activation at CRP-S sites[3]

INTRODUCTION

The cyclic AMP receptor protein, CRP (also called catabolite activator protein, CAP) regulates a global sugar starvation response in three γ-proteobacteria families, the *Enterobacteriaceae*, *Pasteurellaceae*, and *Vibrionaceae* (1). Only *Escherichia coli* CRP has been extensively characterized both structurally and functionally; it thus serves as a model for understanding CRP activity in other bacteria. CRP binds specific sites at gene promoters when activated by its allosteric effector cAMP, and then recruits RNA polymerase through direct protein-protein contacts (reviewed in (2)). CRP binding causes a sharp DNA bend of ~90°; this dramatic deformation of DNA is thought to stimulate transcription both by bringing upstream promoter elements into contact with the polymerase and by facilitating DNA melting (2-4).

CRP homodimers bind 22 bp sequences with 2-fold symmetry (consensus half-site 5'-$A_1A_2A_3T_4G_5T_6G_7A_8T_9C_{10}T_{11}$). DNA site specificity is achieved when a complementary surface shape of base pairs in the DNA major groove allows electrostatic interactions with the helix-turn-helix domain; residues R180, E181, and R185 form hydrogen bonds with bases G_5 and G_7, while Q170, R180, E181, and R185 form non-classical hydrogen bonds with thymine methyl groups at base pairs T:A_4, T:A_6, and A:T_7 (5-7). Additional hydrogen bonds form non-specifically between positively charged protein residues and phosphates in the DNA backbone to reinforce DNA bending. Although CRP does not contact position 6, sequence specificity for T_6 arises through an indirect readout mechanism because a T_6-G_7 base step is favourable to kinking (6, 8).

We have recently described non-canonical CRP sites (CRP-S sites) in *Enterobacteriaceae* and *Pasteurellaceae* competence gene promoters. These differ from canonical (CRP-N) sites in having a C instead of T at position 6. The biological significance of the CRP-S C_6 base is

---

[3] A version of this chapter will be submitted for publication. Cameron A.D.S. and Redfield R.J. (2007) Sxy induces competence by enhancing CRP binding and transcription activation at CRP-S sites.

perplexing because a $T_6 \rightarrow C_6$ substitution reduces CRP-DNA affinity 80-fold (6). In both *E. coli* and *Haemophilus influenzae*, CRP-S promoters are united by dependence on the Sxy protein as well as CRP for expression, raising the possibility that Sxy helps CRP bind DNA at these promoters (Chapters 2 and 3). Thus, Sxy may directly bind CRP-S sites and reduce the free energy requirements of CRP-induced DNA bending. Alternatively, Sxy may not bind DNA but instead act exclusively through contact with CRP to enhance DNA binding or to improve CRP-RNA polymerase (RNAP) contacts.

In addition to contacts formed with transcription factors, RNAP directly recognizes promoter DNA sequence. RNAP's sigma (σ) subunit binds to hexamers located 35 and 10 bp upstream of the transcription start (referred to as the -35 and -10 elements), whereas the two alpha (α) subunits bind A/T runs called UP elements located upstream of -35. Contact between α's C-terminal domain (αCTD) and upstream DNA enhances transcription from 2 to >100-fold at many (if not all) promoters (9, 10). At promoters with weak or absent UP elements, CRP can compensate for low αCTD affinity by making direct and specific contacts with the RNAP subunit(s). At Class I promoters such as P_{lacZYA}, CRP's activating region 1 (AR1) recruits αCTD to DNA immediately downstream of the CRP site to increase RNAP's affinity for the promoter (2, 11, 12). Because all CRP-S sites are predicted to function through a Class I mechanism, Sxy is expected to operate within the Class I framework (Chapter 3). Thus, Sxy may increase promoter occupancy by CRP and leave the AR1 unobstructed for interaction with RNAP, or Sxy may itself directly recruit αCTD.

Traditionally, all *E. coli* CRP sites were thought to belong to a hierarchy in which low-affinity sites are bound only at high concentrations of active CRP or when CRP is displaced from neighbouring high-affinity sites. However, discovery of the functionally distinct CRP-N and CRP-S sub-populations has elaborated this regulon structure. CRP-S sites are also striking because they specifically resist deformation by their cognate transcription factor.

Understanding the physical basis of CRP-DNA binding and CRP-Sxy interactions is key to understanding the mechanisms behind CRP-S promoter activation. We have previously shown that *E. coli* CRP (*Ec*CRP) binds with high affinity to both a CRP-N and a CRP-S promoter from *H. influenzae*, but has very low affinity for its own CRP-S promoters (Chapter 2, Fig. 2.5; Chapter 3, Fig. 3.6). To assess whether *Ec*CRP's DNA-binding properties are representative of CRP from other species, we have purified CRP from *H. influenzae* (*Hi*CRP). Surprisingly,

these two proteins demonstrate different DNA binding affinities and site selectivity: *Ec*CRP has a much higher affinity for DNA, while *Hi*CRP is more discriminating. Nevertheless, two features are shared between *E. coli* and *H. influenzae*: (1) CRP-S sites are low affinity sites for cognate CRP, and (2) even when CRP binds to a CRP-S promoter, it cannot activate significant transcription in the absence of Sxy.

MATERIALS AND METHODS

Strains and culture conditions

H. influenzae cells were cultured at 37°C in brain heart infusion (BHI) supplemented with NAD (2μg/ml) and hemin (10μg/ml), including novobiocin (2.5μg/ml), kanamycin (7μg/ml), or cholramphenicol (2μg/ml) when required. *H. influenzae* cells were transformed with chromosomal or plasmid DNA as previously described (13). *E. coli* DH5α was made chemically competent with RbCl and transformed as previously described (14).

Protein purification and bandshifts

H. influenzae crp coding sequence was cloned under P_{lac} control in the His-tag vector pQE30 (Qiagen); *E. coli crp* was cloned by Peekhaus and Conway (15) using the same method. His-tagged proteins were expressed and purified as previously described in Chapter 3. Native *Hi*CRP was isolated from a *H. influenzae cya-* mutant using the technique described for isolation of *Ec*CRP in Chapter 2. Reaction conditions for bandshift assays are also described in Chapter 3. For each protein dilution series, a line fit to the data was used to calculate the K_d for that assay. K_d values from two or more independent dilution series were averaged, and these values are reported in the text. Fresh bait DNA and freshly thawed protein were used in each assay.

Cloning and site-directed mutagenesis

The *H. influenzae pilABCD* operon promoter (P_{pil}) was cloned as follows. The chromosomal region (coordinates 333193-335531) containing *ampD*, *pilA*, and the N-terminal half of *pilB* was PCR amplified and cloned in pGEM-T Easy (Promega). An AccI digest was used to excise a fragment containing P_{pil}, *pilA*, *pilB* and 20 bp of multiple-cloning site. This fragment was cloned in the AccI site in the *H. influenzae* cloning vector pSU20 to generate plasmid p*pilA*. pSU20 contains a *lacZ* promoter (P_{lac}) adjacent to the multiple-cloning site that is constitutively

expressed in *H. influenzae*. To prevent P_{lac} from interfering with P_{pil} induction, P_{lac} was removed by a XmnI and XhoI double digest and vector DNA was purified on an agarose gel. The sticky end generated by XhoI was filled using the Klenow fragment of DNA Polymerase I and this was ligated to the XmnI-cut blunt end to generate plasmid p*pilA*::*Plac*(-).

P_{pil} in p*pilA*::*Plac*(-) and P_{lac} in pSU20 were mutated to P_{pil}-*N* and P_{lac}-*S*, respectively, using Stratagene's QuickChange Site-Directed Mutagenesis kit according to the manufacturer's instructions. Mutagenesis primers: PpilF 5'-ATTGACCGCACTTTTTCTGTGATCCTGATCACAAAAAAAAGGAAAAATGTAT; PpilR 5'-ATACATTTTTCCTTTTTTTTGTGATCAGGATCACAGAAAAAGTGCGGTCAAT.

The pSU20 and pASKA vectors have compatible origins of replication, but both confer chloramphenicol resistance. In order to propagate either P_{lac} or P_{lac}-*S* in the same cells as pASKA plasmids, both promoters were excised from pSU20 using an XmnI and XhoI double digest. Each was cloned into XmnI/XhoI digested pSU40, which confers kanamycin resistance, to generate plasmids pSU40::*Plac* and pSU40::*Plac-S*.

Real time (quantitative) PCR

Real time PCR was conducted as described in Chapter 3. Because we only wanted to detect transcripts originating from the plasmid-borne promoters that we engineered, primers were designed to amplify only cDNA from plasmid-encoded transcripts. To measure P_{pil} and P_{pil}-*N* activity, PCR primers were designed to flank the junction of *pilB* and the multiple-cloning site in p*pilA*. Primers: pilB-RTF 5'-TCTGCCTTACAAAAAAATGCCTCTG; ppilA-RTR 5'-GGGGATCCTCTAGAGTCGACCTGC. This primer set did not generate amplicons when chromosomal DNA was used as template, confirming that it targets only plasmid-encoded P_{pil} transcripts.

RESULTS

E. coli and *H. influenzae* CRP display different binding-site affinities and specificities

CRP alone cannot activate CRP-S promoters (Chapter 2). As a first step to understand why, we tested CRP binding to a collection of sequences containing either CRP-N or CRP-S promoters. *Ec*CRP and *Hi*CRP were purified in native form on cAMP-affinity columns or were His-tagged for purification on nickel-affinity columns. Because of their consistently high yields and purity,

His-tagged proteins were used in electrophoretic mobility (bandshift) assays to quantify DNA binding. Four different bait DNAs were tested (CRP sites shown in Fig. 4.1A): two CRP-N promoters, *H. influenzae* P_{mglBAC} (P_{mgl}) and the archetypal CRP-regulated *E. coli* P_{lacZYA} (P_{lac}), and two *H. influenzae* CRP-S promoters, $P_{comA\text{-}E}$ (P_{com}) and $P_{pilABCD}$ (P_{pil}). CRP was added to binding reactions and allowed time to equilibrate; CRP-DNA dissociation constants (K_d) were calculated as the protein concentration at which half of the bait DNA was bound.

*Ec*CRP demonstrated the greatest affinity for P_{mgl} (K_d 3 ±1 nM) (Fig. 4.1B), consistent with the perfect core $T_4G_5T_6G_7A_8$ sequence in both halves of the P_{mgl} CRP site. Our measurement of K_d 7.5 nM (quantified from one dilution series) for P_{lac} compares well with 10.8 nM measured by Ebright *et al.* (16) using a filter-binding assay. *Ec*CRP showed lower affinity for the two CRP-S sites: P_{com} K_d 45 nM (quantified from one dilution series) and P_{pil} K_d 70±20 nM.

*Hi*CRP also showed the greatest affinity for P_{mgl} (K_d 70±30 nM), but this is 25-fold less than *Ec*CRP's affinity for the same site (Fig. 4.1B). Surprisingly, *Hi*CRP did not bind P_{lac}, P_{com}, or P_{pil} until protein concentrations were so high as to elicit non-specific DNA binding (>2000 nM; not shown. These assays were repeated 2 or more times). To ensure that this inability to bind CRP-S promoters was not due to interference by the N-terminal histidine tag, DNA binding by native *Hi*CRP was assayed. Native *Hi*CRP bound to P_{mgl} but not P_{com} (Fig. 4.1C), as seen with His-tagged protein.

Together, these results show that *Ec*CRP and *Hi*CRP differ in two ways. First, *Ec*CRP's greater affinity for all CRP sites tested shows it to be a stronger DNA binding protein. Second, *Ec*CRP has only a 25-fold greater affinity for P_{mgl} CRP-N over the P_{com} and P_{pil} CRP-S sites whereas *Hi*CRP has a greater than 1000-fold preference for P_{mgl} over P_{com} and P_{pil}. The inability of *Hi*CRP to bind P_{lac}, P_{com}, or P_{pil} *in vitro* suggests that this protein has high selectivity for the perfect core $T_4G_5T_6G_7A_8$ sequence.

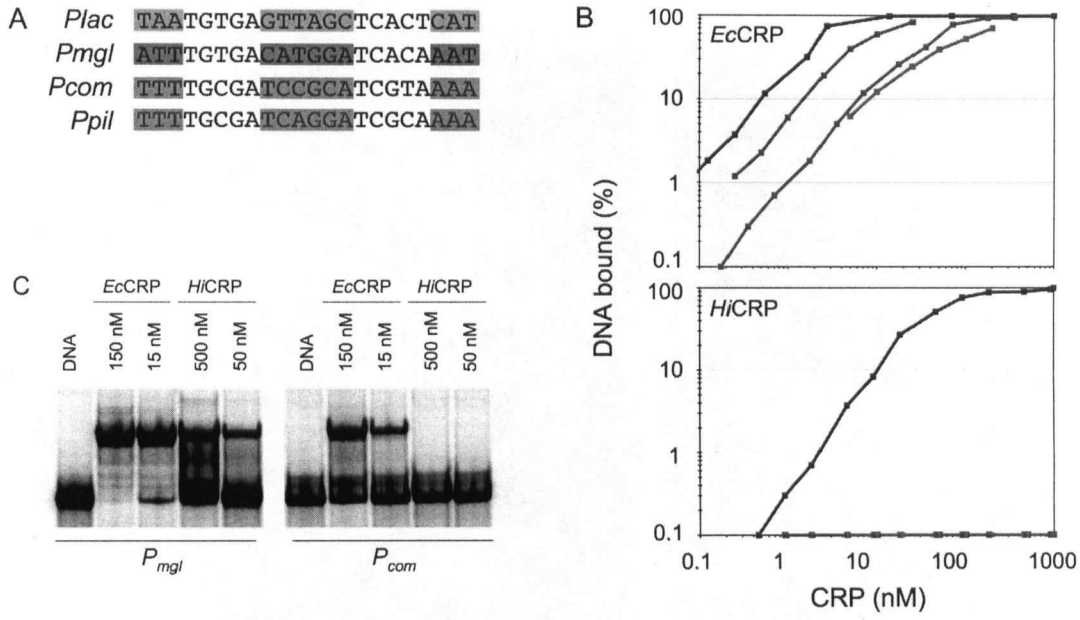


Figure 4.1 CRP binding to *E. coli* and *H. influenzae* promoters.
A. CRP site sequences. **B.** CRP-DNA binding quantified using bandshift assays. **C**. Bandshift assay to test binding of native proteins to P_{mgl} and P_{com}.

Three-dimensional mapping of *Hi*CRP

Multiple crystal structures of *Ec*CRP binding to DNA have been solved (6, 7, 17), allowing us to map *Hi*CRP residues on the *Ec*CRP tertiary structure using SWISS-MODEL (18). We examined the predicted *Hi*CRP structure for residues that may disrupt DNA binding, either by reshaping the DNA-binding domain or by sterically interfering with protein-DNA interactions that are known to occur in *E. coli*.

*Ec*CRP's 50 C-terminal amino acids contain the helix-turn-helix DNA-binding domain and additional residues that contact DNA (highlighted in Figure 3.8). This region is 92% identical and 96% similar between *Ec*CRP and *Hi*CRP, permitting construction of a very high-confidence *Hi*CRP structure modeled on *Ec*CRP bound to a synthetic CRP-N site (Protein data base file 1ZRC, solved by (6). Figure 4.2 shows the two C-terminal regions of a *Hi*CRP dimer viewed from four vantage points. Three of the four non-conserved amino acids are highlighted red (the *Ec*CRP template lacks the two terminal amino acids, one of which is not conserved in *Hi*CRP, but these terminal amino acids are not known to play a role in DNA binding).

In this model, none of the three non-conserved amino acids are positioned close to DNA, nor are they exposed on the face of the protein that contacts DNA. However, two are adjacent to amino

acids that make DNA contacts, raising the possibility that they slightly modify the surface shape of CRP and so upset protein-DNA contacts. We used a suite of algorithms designed to validate predicted protein structures (Procheck, Prove, WhatIf; available at http://biotech.ebi.ac.uk/) to examine the potential of non-conserved amino acids to alter *Hi*CRP's shape. The predicted *Hi*CRP model was found to be completely congruous with the *Ec*CRP structure. Thus, this structural analysis does not implicate any residues in *Hi*CRP's C-terminal domain in reducing the protein's affinity for DNA.

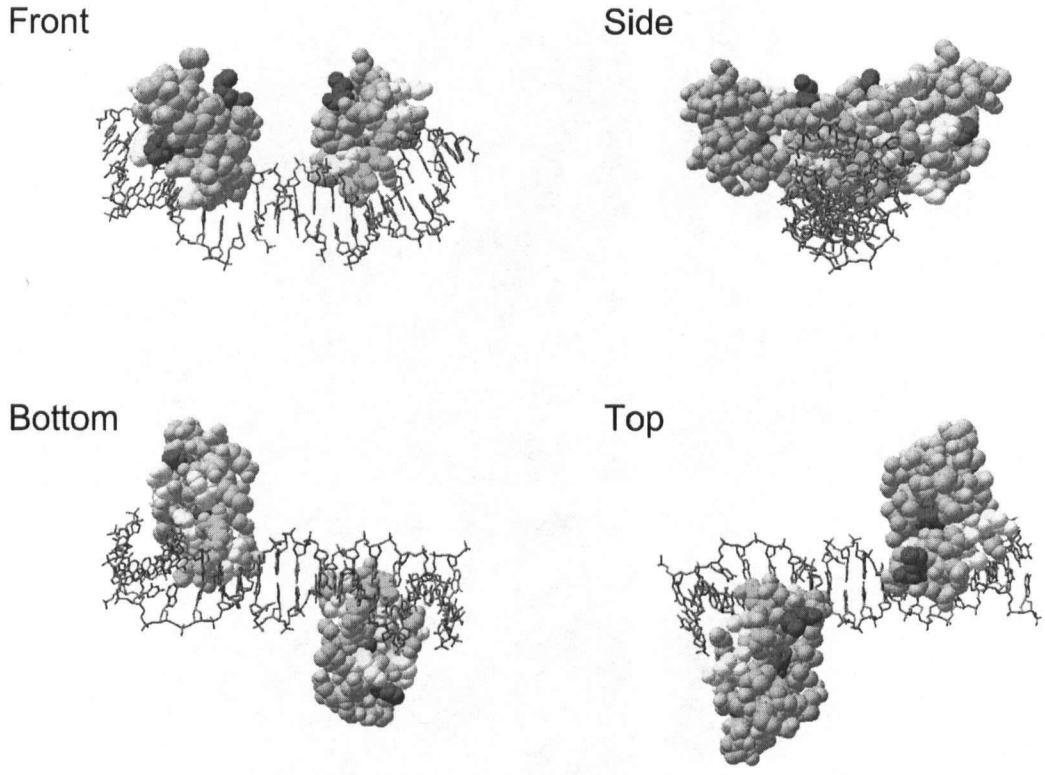


Figure 4.2 Predicted tertiary structure of *Hi*CRP C-terminal domains bound to a CRP-N site.
Amino acids highlighted as follows: Green, amino acids making base contacts; Yellow, amino acids making phosphate contacts; Red, amino acids that differ between *Ec*CRP and *Hi*CRP; Blue, all amino acids shared between *Ec*CRP and *Hi*CRP. The DNA is grey (22 bp site).

The CRP-S C_6 base prevents *Hi*CRP binding

Because the C_6-G_7 base step characteristic of CRP-S sites inhibits CRP-induced DNA kinking (6, 19), conversion of C_6-G_7 to a canonical T_6-G_7 step in each half of the P_{pil} promoter was predicted to enable *Hi*CRP binding. Site-directed mutagenesis was used to convert C_6 to T_6 in both halves of P_{pil}, generating P_{pil}-*N* (Fig. 4.3A). Figure 4.3B shows that *Hi*CRP binds

specifically to P_{pil}-*N* (blue line) in contrast to the complete absence of binding to P_{pil} (red line; data from Fig. 4.1), confirming that *Hi*CRP is highly selective for the core sequence $T_4G_5T_6G_7A_8$. Figure 4.3C shows that *Ec*CRP has a 20-fold greater affinity for P_{pil}-*N* over P_{pil}, comparable to its 50-fold preference for P_{mgl} over P_{pil}. Thus, the C_6 base is sufficient to explain *Hi*CRP's inability to bind CRP-S sites, and indicates that sequence flanking the $T_4G_5T_6G_7A_8$ core does not prevent *Hi*CRP from binding to a CRP-S site. In addition, the C_6 base accounts for *Ec*CRP's preference for P_{mgl} over P_{com} and P_{pil}.

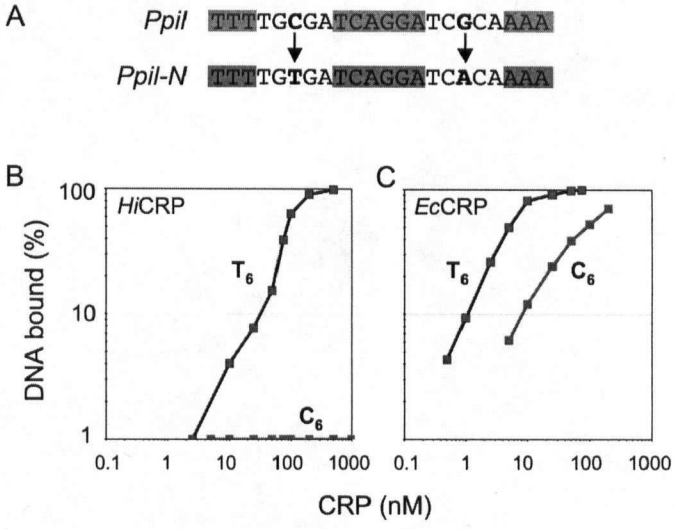


Figure 4.3 Mutagenesis of the *pilA-D* operon CRP-S site to resemble a CRP-N site.
A. DNA sequences of the wildtype (P_{pil}) and mutated (P_{pil}-*N*) site. **B+C.** CRP-DNA binding quantified using bandshift assays with *Hi*CRP or *Ec*CRP, respectively.

The C_6 to T_6 mutation is highly deleterious to promoter activity

The inability of *Hi*CRP to bind P_{com} and P_{pil} *in vitro* suggested that Sxy is required for transcription because it helps *Hi*CRP bind to these promoters. Because *Hi*CRP can bind P_{pil}-*N*, we hypothesized that transcription would initiate from P_{pil}-*N* in the absence of Sxy; this was tested using real-time PCR to quantify promoter activity in *H. influenzae*. P_{pil} was induced 120-fold in wildtype cells but was not induced in the absence of *sxy* (Fig. 4.4), consistent with the expression levels measured using microarrays in Chapter 2. We expected the mutated promoter to stimulate equally high levels of transcription. Surprisingly, P_{pil}-*N* was induced only 8-fold in wildtype cells, indicating that conversion of C_6 to T_6 in both half-sites was highly deleterious to promoter activity even in the presence of Sxy. Most notably, P_{pil}-*N* was induced 3-fold in the absence of *sxy*, unlike its completely *sxy*-dependent parent P_{pil}.

Considering *Hi*CRP's ability to bind P_{pil}-*N in vitro*, these *in vivo* results suggest that *Hi*CRP binding alone is insufficient to stimulate wildtype transcription levels. Moreover, increasing promoter occupancy by *Hi*CRP has reduced Sxy's influence; Sxy has been demoted from an essential transcription factor to an accessory protein that contributes only a 2-fold increase in transcription at P_{pil}-*N*.

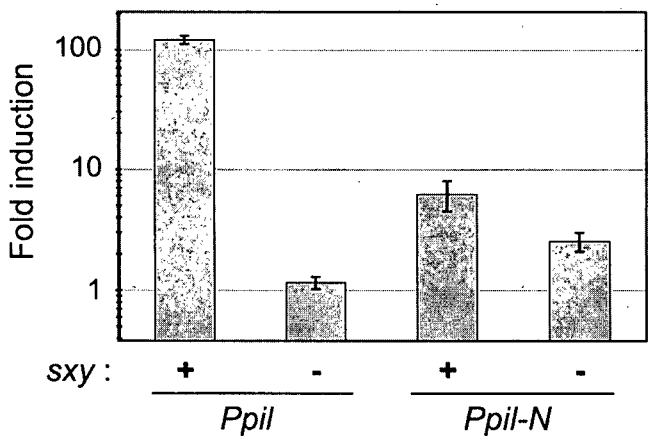


Figure 4.4 Real-time PCR quantification of *Ppil* and *Ppil-N* activity in *sxy+* and *sxy- H. influenzae* cells.
Gene expression was measured 0 and 60 minutes after transfer of cells to the strong competence inducing medium MIV. Fold induction after 60 minutes is plotted relative to *pil* operon expression levels in the uninduced state at 0 minutes; the mean and range of two independent cultures are plotted on a log scale.

*Ec*CRP requires Sxy to activate *H. influenzae* competence genes

*Ec*CRP's relatively high affinity for *H. influenzae's* P_{com} and P_{pil} provided an alternate means to test whether CRP alone can activate transcription when it binds a CRP-S promoter in the absence of Sxy. The plasmid pNX15, which carries the *E. coli crp* gene (20), was cloned in *H. influenzae crp-* or *sxy-* mutants. Transcription from P_{com} was measured after 60 minutes in MIV, the time at which competence genes are maximally expressed (Chapter 2). Transformation frequency provides a more sensitive assay of competence gene induction, so it was measured after 90 minutes in MIV.

Figure 4.5 shows that *Ec*CRP complemented the *Hi*CRP null mutant to restore P_{com} expression and natural transformation, as reported previously (20). However, even though *Ec*CRP binds P_{com} with high affinity *in vitro*, *Ec*CRP did not induce P_{com} or restore transformability in *sxy-*

cells. These results provide further evidence that Sxy is required for transcription activation, not just for CRP binding.

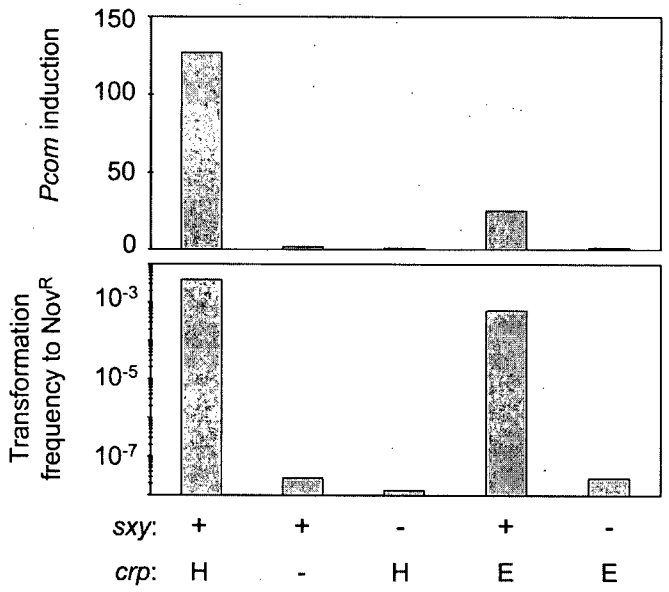


Figure 4.5 *Ec*CRP complementation of *Pcom* induction and natural transformation in *H. influenzae sxy-* and *crp-* cells.
H = *H. influenzae crp;* E = *E. coli crp.* Gene expression and transformation were measured after 60 and 90 minutes in MIV, respectively. P_{com} induction is presented as fold induction relative to *com* operon expression levels in *sxy-* cells with *Hi*CRP.

Conserved motifs in *H. influenzae* CRP-S promoters may be UP elements

Previous searches for putative Sxy binding sites in *H. influenzae* have not identified any conserved motifs other than CRP-S sites (Chapter 2). However, Sxy has a strong influence on transcription activation at P_{pil}, but not P_{mgl} (Figure 3.7) or P_{lac} (not shown) suggesting that CRP-S promoters contain specialized features in addition to the $T_4G_5C_6G_7A_8$ motif. To detect conserved elements in *H. influenzae*'s CRP-S regulon promoters, we generated a sequence logo from alignment of the 13 DNA sequences at their CRP sites. This revealed a significant overrepresentation of evenly spaced A/T runs at positions -79, -90, and -102 (Fig. 4.6A). The A/T runs are spaced such that the minor groove is on the same face of the DNA as CRP and RNAP; this is significant because αCTD binds the DNA minor groove. The sequence logo downstream of CRP-S sites exhibits overrepresentation of sequences resembling the *E. coli* σ70 -35 and -10 binding sites. UP elements are usually found between the σ70 -35 binding site and the CRP site, but these CRP-S promoters have no sequence conservation in this region. This

raises the possibility that Sxy recruits αCTD to UP elements upstream of the CRP site in P_{pil} (this is illustrated in Figure 4.6B).

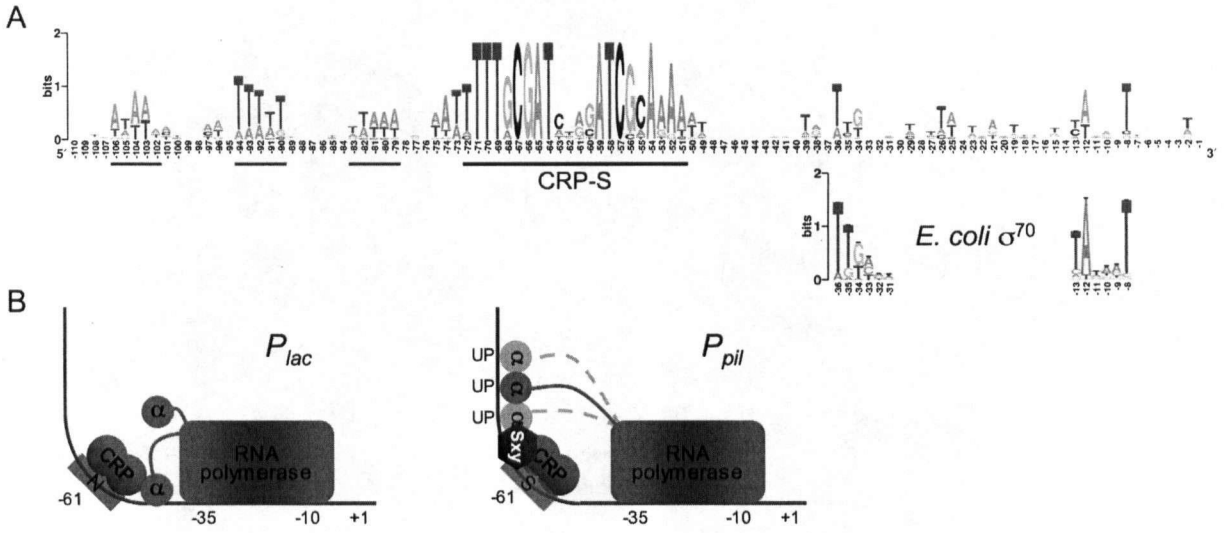


Figure 4.6 Putative UP elements in competence gene promoters.
A. Sequence logo generated from alignment of *H. influenzae*'s 13 CRP-S promoters. Numbering indicates the average distance of CRP-S sites from predicted transcription start sites (presented in Fig. 2.3); putative UP elements are underlined blue. Sequence logos generated from alignment of 401 *E. coli* σ70 binding sites (copied from (27)) facilitate comparison with similar motifs in *H. influenzae* CRP-S promoters. **B.** Illustration of αCTD-DNA contact at P_{lac} and the proposed contacts at P_{pil}.

DISCUSSION

This study of the molecular mechanisms regulating competence genes in *E. coli* and *H. influenzae* is both the first detailed analysis of CRP binding to competence gene promoters and the first analysis of CRP from a member of the *Pasteurellaceae*. Both *Ec*CRP and *Hi*CRP bind specifically to DNA sites containing the core CRP-N sequence $T_4G_5T_6G_7A_8$. On the other hand, *Hi*CRP cannot bind its own competence gene promoters containing the CRP-S sequence $T_4G_5C_6G_7A_8$, and *Ec*CRP demonstrates very low affinity for its cognate CRP-S promoters (Figure 3.6). Thus a hallmark of *Enterobacteriaceae* and *Pasteurellaceae* CRP-S is that they are specific but low-affinity CRP sites.

Low affinity for CRP-S sites is achieved in a species-specific fashion that corresponds to CRP's affinity for DNA. In *H. influenzae*, CRP-S sequences are all strong matches to the CRP-binding site consensus, but always include a stiff C6-G7 base step that is expected to preclude DNA bending and binding by *Hi*CRP alone (23). In *E. coli*, CRP-S sites vary from the CRP-binding site consensus at many positions and always have either a C6-G7 or G6-G7 base step (1).

Therefore, CRP-S sites appear to prevent CRP occupancy (and subsequently transcriptional activation) in the absence of Sxy. Under this model, Sxy either first binds promoter DNA or binds CRP in solution, and then assists CRP binding at CRP-S sites to activate transcription.

Why does conversion of P_{pil} to P_{pil}-*N* reduce promoter activity in wildtype *H. influenzae* cells? The simplest explanation is that Sxy-CRP more effectively recruits RNA polymerase to CRP-S promoters than CRP does to CRP-N promoters. We posit that because CRP can bind P_{pil}-*N* unassisted, it out-competes Sxy-CRP for DNA binding, consequently reducing promoter activity. Conversely, Sxy has no apparent effect on P_{lac} induction (not shown) or P_{mgl} (Figure 3.7), suggesting that it does not interfere with normal CRP activity at CRP-N promoters.

We have repeatedly observed that plasmid-borne *sxy* accumulates mutations unless its expression is tightly repressed, indicating that constitutive or over-expression of Sxy is toxic to cells. Cells may avoid toxic effects by maintaining low concentrations of Sxy, in which case Sxy-CRP complexes would represent only a small fraction of total cellular CRP. Low levels of Sxy-CRP should be sufficient to activate the small fraction of CRP-S among total genomic CRP sites, while at the same time preventing Sxy from sequestering too much CRP.

It has been shown that adding artificial UP elements between the CRP and RNAP-binding sites increases promoter strength up to 15-fold in *E. coli* (24, 25). The presence of putative UP elements upstream of *H. influenzae* CRP-S sites offers an intriguing explanation for the strength of transcription activation at these promoters. Lee and coworkers (26) have shown that when CRP binds at positions -60.5 or -69.5, an αCTD subunit can be recruited to positions -80 or -91 respectively, indicating that upstream DNA is accessible to RNAP. These coordinates correspond to the location of most *H. influenzae* CRP-S sites and of putative UP elements in Figure 4.6A. Deletion analysis of sequences upstream of CRP-S sites will resolve the importance of these A/T runs in promoter activity. Also, if engineering P_{pil}-*N* to have UP elements at the more common positions of -52 and -42 removes the need for Sxy, this will implicate Sxy as a mediator of αCTD binding to upstream sequence in CRP-S promoters. Further, A/T runs can be introduced upstream of the P_{lac} CRP site to test whether Sxy enhances the expression of this promoter.

H. influenzae CRP shares 78% identity with its *E. coli* ortholog, and all residues that contact DNA are conserved in both proteins. This level of sequence identity strongly suggested that both proteins would exhibit very similar DNA-binding properties. However, the results

presented here show that *H. influenzae* CRP has a much lower affinity for DNA. These unexpected properties are unlikely to result from differences in cAMP binding characteristics; all residues that contact cAMP in *E. coli* are conserved in *H. influenzae*, and cAMP was in excess in CRP-DNA binding assays. Another possibility is that *H. influenzae* CRP forms fewer direct contacts with DNA, but our modeling of *Hi*CRP tertiary structure supports the prediction that *Hi*CRP can form all bonds found with *Ec*CRP. To test this, the 4 amino acid differences from *Hi*CRP should be introduced into the *Ec*CRP DNA-binding domain. If these differences do affect protein-DNA contacts, *Ec*CRP will demonstrate reduced affinity for CRP-S sites and P_{lac}. It is also possible that *Hi*CRP forms less-stable dimers than the *E. coli* protein. Given the very low concentrations at which CRP is active in DNA binding (1-100 nM, Figure 4.1B), comparing dimerization strength will require a very sensitive assay.

Chen *et al.*, (6) have shown that *Ec*CRP forms a hydrogen bond between E181 and C_6 at a CRP-S site, which does not form at a CRP-N site. However, this additional bond cannot compensate for the C6-G7 step's resistance to deformation, and CRP continues to strongly favour binding to the flexible $T_4G_5T_6G_7A_8$ sequence. Chen et al., (6) unexpectedly discovered that the substitution E181A causes CRP to preferentially bind $T_4G_5C_6G_7A_8$ over the canonical $T_4G_5T_6G_7A_8$ sequence, even though overall the protein has greatly decreased affinity for binding sites. All γ-proteobacteria CRP molecules have E181 (aligned in Figure 3.8), thus CRP-S are second-rate sites in all members of this group.

REFERENCES

1. Cameron,A.D. and Redfield,R.J. (2006) Non-canonical CRP sites control competence regulons in Escherichia coli and many other gamma-proteobacteria. *Nucleic Acids Res.*, **34**, 6001-6014.
2. 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.
3. Eichenberger,P., Dethiollaz,S., Buc,H. and Geiselmann,J. (1997) Structural kinetics of transcription activation at the malT promoter of Escherichia coli by UV laser footprinting. *Proc Natl Acad Sci U S A,* **94**, 9022-9027.
4. Coulombe,B. and Burton,Z.F. (1999) DNA bending and wrapping around RNA polymerase: a "revolutionary" model describing transcriptional mechanisms. *Microbiol Mol Biol Rev,* **63**, 457-478.
5. Mandel-Gutfreund,Y., Margalit,H., Jernigan,R.L. and Zhurkin,V.B. (1998) A role for CH.O interactions in protein-DNA recognition. *J. Mol. Biol.*, **277**, 1129-1140.
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. 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.
8. 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.
9. Ross,W., Aiyar,S.E., Salomon,J. and Gourse,R.L. (1998) Escherichia coli promoters with UP elements of different strengths: modular structure of bacterial promoters. *J. Bacteriol.*, **180**, 5375-5383.
10. Ross,W. and Gourse,R.L. (2005) Sequence-independent upstream DNA-alphaCTD interactions strongly stimulate Escherichia coli RNA polymerase-lacUV5 promoter association. *Proc Natl Acad Sci U S A,* **102**, 291-296.
11. Kolb,A., Igarashi,K., Ishihama,A., Lavigne,M., Buckle,M. and Buc,H. (1993) E. coli RNA polymerase, deleted in the C-terminal part of its alpha-subunit, interacts differently with the cAMP-CRP complex at the lacP1 and at the galP1 promoter. *Nucleic Acids Res.*, **21**, 319-326.
12. Busby,S. and Ebright,R.H. (1999) Transcription activation by catabolite activator protein (CAP). *J. Mol. Biol.*, **293**, 199-213.
13. Poje,G. and Redfield,R.J. (2003) General methods for culturing Haemophilus influenzae. *Methods Mol Med,* **71**, 51-56.

14. Ausubel,F.M. (1995) Current protocols in molecular biology. J. Wiley & Sons, Inc, Brooklyn, NY.

15. 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.

16. Ebright,R.H., Ebright,Y.W. and Gunasekera,A. (1989) Consensus DNA site for the Escherichia coli catabolite gene activator protein (CAP): CAP exhibits a 450-fold higher affinity for the consensus DNA site than for the E. coli lac DNA site. *Nucleic Acids Res.*, **17**, 10295-10305.

17. 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.

18. Schwede,T., Kopp,J., Guex,N. and Peitsch,M.C. (2003) SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res.*, **31**, 3381-3385.

19. 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.

20. Chandler,M.S. (1992) The gene encoding cAMP receptor protein is required for competence development in Haemophilus influenzae Rd. *Proc Natl Acad Sci U S A*, **89**, 1626-1630.

21. Gourse,R.L., Ross,W. and Gaal,T. (2000) UPs and downs in bacterial transcription initiation: the role of the alpha subunit of RNA polymerase in promoter recognition. *Mol. Microbiol.*, **37**, 687-695.

22. Ross,W., Ernst,A. and Gourse,R.L. (2001) Fine structure of E. coli RNA polymerase-promoter interactions: alpha subunit binding to the UP element minor groove. *Genes Dev.*, **15**, 491-506.

23. 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.

24. Czarniecki,D., Noel,R.J.J. and Reznikoff,W.S. (1997) The -45 region of the Escherichia coli lac promoter: CAP-dependent and CAP-independent transcription. *J. Bacteriol.*, **179**, 423-429.

25. Noel,R.J.J. and Reznikoff,W.S. (1998) CAP, the -45 region, and RNA polymerase: three partners in transcription initiation at lacP1 in Escherichia coli. *J. Mol. Biol.*, **282**, 495-504.

26. Lee,D.J., Busby,S.J. and Lloyd,G.S. (2003) Exploitation of a chemical nuclease to investigate the location and orientation of the Escherichia coli RNA polymerase alpha subunit C-terminal domains at simple promoters that are activated by cyclic AMP receptor protein. *J. Biol. Chem.*, **278**, 52944-52952.

27. Shultzaberger,R.K., Chen,Z., Lewis,K.A., and Schneider,T.D. (2007) Anatomy of Escherichia coli sigma70 promoters. *Nucleic Acids Res.*, **35**, 771-788.

CHAPTER FIVE

RNA secondary structure regulates *sxy* expression and competence development in *Haemophilus influenzae*[4]

INTRODUCTION

Natural competence, the ability to take up DNA molecules directly from the environment, is tightly regulated in most bacteria, indicating that the costs and benefits of DNA uptake depend on changes in the extracellular and intracellular environments. Because the mechanisms regulating competence evolved to allow cells to track these changes, understanding the mechanisms provides a window on the importance of DNA uptake to the cell.

Bacteria in the families *Pasteurellaceae*, *Enterobacteraceae* and *Vibrionaceae* appear to share a common regulatory mechanism, with competence genes organized in a regulon whose transcription is controlled by two activator proteins, Sxy (also known as TfoX) and CRP (also known as CAP) (1). Conditions that induce competence have been well-studied only in *Haemophilus influenzae* (*Pasteurellaceae*). *H. influenzae* becomes moderately competent as growth slows during late log phase in rich medium, and becomes maximally competent when log phase cells are transferred to the defined starvation medium MIV (2).

The *sxy* gene was first identified and named as the site of the *H. influenzae* mutation *sxy-1*, which causes greatly increased competence (hypercompetence) during growth in rich medium (3). The competence regulon it controls contains 25 genes (13 transcription units), many known to contribute directly to DNA uptake (2). Cells carrying a *sxy* knockout cannot induce the competence regulon; conversely, overexpression of *sxy* from multi-copy plasmids induces competence under what are normally non-inducing conditions in both *H. influenzae* and *V. cholerae* (4, 5). Consistent with its role in competence induction, *H. influenzae sxy* mRNA levels rise when cells experience competence-inducing conditions (2).

Unlike Sxy, CRP is a global regulator. It activates a broad array of genes united by their roles in obtaining or utilizing alternative carbon or energy sources, or in sparing the wasteful use of the

---

[4] A version of this chapter has been submitted for publication. Cameron A.D.S., Volar M., Bannister L.A., and Redfield R.J. (2007) RNA secondary structure regulates *sxy* expression and competence development in *Haemophilus influenzae*. *Submitted to Molecular Microbiology.*

preferred sources; its action has been very well studied in *Escherichia coli*. When activated by its allosteric effector cAMP, CRP binds specifically to 22bp sites in promoters where it stimulates transcription. This stimulation is triggered by a rise in cAMP that occurs when preferred sugars are unavailable for transport by the phosphotransferase system (6, 7). Competence regulon promoters contain a novel type of CRP site (CRP-S sites, previously called competence regulatory elements), whose sequences differ at critical positions from the canonical CRP sites previously characterized in *E. coli* (CRP-N sites) (1). Because promoters that depend on both CRP and Sxy for transcription activation have only the single CRP-S factor binding site, the CRP-S motif is thought to be the sequence determinant of Sxy activity. Sxy's lack of a distinct DNA binding site distinguishes it from all of the previously characterized co-regulators of CRP.

The original *sxy-1* mutation causes only a very conservative change in the Sxy protein sequence ($Val_{19}Ile$), and it was proposed to cause hypercompetence by increasing the amount of Sxy rather than by changing the nature of Sxy's action (4). Here we report the isolation and characterization of additional hypercompetence-causing point mutations in *sxy*. We show that all of the *sxy* hypercompetence mutations act by increasing *sxy* expression, and that these effects arise by destabilization of an mRNA secondary structure that normally limits *sxy* expression in rich medium. In maximal competence-inducing conditions, CRP induces *sxy*; together CRP and Sxy then induce the genes of the CRP-S regulon.

MATERIALS AND METHODS

Strains, plasmids, and DNA

Bacterial strains and plasmids used in this study are listed in Table 5.1.

Table 5.1 Strains used in this work

Strain or plasmid	Relevant genotype	Source or reference
H. influenzae		
KW20	Wildtype; Sequenced strain	(18)
MAP7	KW20 Nov^r	(19)
RR648	KW20 *sxy*::Kan^r	(4)
RR699	KW20 *sxy-1*	(4)
RR700	KW20 *sxy-2*	This study
RR723	KW20 *sxy-3*	This study
RR724	KW20 *sxy-4*	This study
RR844	KW20 sxy_{89}::*lacZKan*^{r}(operon fusion)	This study
RR845	KW20 sxy_{89}::*lacZKan*^{r}(protein fusion)	This study

RR846	KW20 *sxy*$_{11}$::*lacZKan*^{r}(operon fusion)	This study
RR847	KW20 *sxy*$_{11}$::*lacZKan*^{r}(protein fusion)	This study
RR850	KW20 *sxy-5*	This study
RR852	KW20 *sxy-6*	This study
RR854	KW20 *sxy-7*	This study
E. coli		
DH5α	*F80lacZ Δ(lacIZYA-argF) endA1*	
JM109	*endA1, recA1*	Promega
M15	*lacZ*, pREP4	Qiagen
RR1128	M15 pQE*sxy*	This study
Plasmids		
pDJM90	UTR and 5'-half of *sxy* ORF	(4)
pLZK80	*lacZKan*^{r} operon fusion cassette	G. Barcak
pLZK81	*lacZKan*^{r} protein fusion cassette	G. Barcak
pLBSF1	*sxy*$_{89}$::*lacZKan*^{r}(operon fusion)	This study
pLBSF2	*sxy*$_{89}$::*lacZKan*^{r}(protein fusion)	This study
pLBSF3	*sxy*$_{11}$::*lacZKan*^{r}(operon fusion)	This study
pLBSF4	*sxy*$_{11}$::*lacZKan*^{r}(protein fusion)	This study
pGEM-7Zf-	T7 promoter	Promega
pGEM*sxy*	*sxy* ORF and UTR cloned in pGEM-7Zf-	This study
pGEM*sxy-1*	*sxy-1* ORF and UTR cloned in pGEM-7Zf-	This study
pGEM*sxy-7*	*sxy-7* ORF and UTR cloned in pGEM-7Zf-	This study
pREP4	*lacI*q	Qiagen
pQE*sxy*	*sxy* ORF cloned in pQE-30UA (Qiagen)	This study

Nov^r, novobiocin resistance; Kan^r, kanamycin resistance; Cam^r, chloramphenicol resistance.

Culture conditions and transformation assays

H. influenzae cells were cultured at 37°C in brain heart infusion (BHI) supplemented with NAD (2 μg/ml) and hemin (10 μg/ml), including novobiocin (2.5 μg/ml), kanamycin (7 μg/ml), or cholramphenicol (2 μg/ml) when required. Competence was induced by transferring cells to MIV medium, as previously described (8). *E. coli* cells were grown in Lauria-Bertani (LB) medium with kanamycin (25 μg/ml) and ampicillin (100 μg/ml) when required.

H. influenzae cells were transformed with chromosomal or plasmid DNA as previously described (9), using 1μg/ml of MAP7 DNA for 15 minutes, with subsequent selection for novobiocin resistance. *E. coli* cells were made chemically competent with RbCl and transformed with plasmids as previously described (10).

Site-directed mutagenesis

The 1.8 kb *Eco*RI-*Bam*HI fragment of pDJM90 (*sxy*) was cloned into the *Eco*RI-*Bam*HI site of pAlter-1 (Promega) to create the plasmid pAltersxy. Site-directed mutagenesis was carried out using the Altered Sites II (Promega), following the manufacturers' protocol. Sequencing was

used to confirm mutations, and the sequenced region between *Sca*I-*Cla*I sites was subcloned into pDJM90 to ensure that the plasmid inserts contained no additional, undesirable mutations.

Generation of polyclonal anti-Sxy antibodies

The *sxy* coding sequence was cloned under *lac* promoter control in the His-tag vector pQE30-UA (Qiagen) in *E. coli* M15, and *sxy* expression was induced at OD_{600} 0.6 with 1mM IPTG. Cells were harvested after 4.5hr by centrifugation and the pellet was frozen overnight at -20°. Sxy invariably formed inclusion bodies in expression cultures, even when induced with low concentrations of IPTG and at 30°. Sxy was denatured and purified as follows: the frozen cell pellet was resuspended in lysis buffer (100 mM $NaPO_4$, 10 mM Tris HCl, 6 M guanidine HCl, pH 8.0), then cells were incubated 1hr at 30° with shaking followed by brief, gentle vortexing until the solution was translucent. Cellular debris was removed by centrifugation at 10,000g for 25 minutes and the supernatant was incubated with nickel-nitriloacetic acid agarose beads for 1 hr at 4° with gentle rocking. Agarose beads were loaded in a column and washed twice with 12 column volumes of wash buffer (100 mM $NaPO_4$, 10 mM Tris HCl, 8 M urea, pH 6.3), and protein was eluted in two steps, each using two volumes of wash buffer, at pH 5.9 and pH 4.5 respectively. Eluted fractions were pooled and concentrated by precipitation with 10% TCA. Residual TCA was removed with cold 100% ethanol, and protein was dried and resuspended in sample buffer (45 mM Tris HCl pH 7.5, 10% glycerol, 1% SDS, 50 mM DTT, 0.01% bromophenol blue). Protein was then run on a 15% polyacrylamide SDS gel, and the section of gel containing Sxy (MW 25 kDa) was excised and macerated by repeated passage through a small-bore syringe. Protein was eluted overnight in water, and then concentrated by TCA precipitation. Dried protein was resuspended in phosphate-buffered saline and purity was assessed with SDS-PAGE and quantified using the Bradford assay. Protein was emulsified in incomplete Freund's adjuvant (250μg/ml) for injection in rabbits. Blood serum was collected 10 days after booster shots and stored at -20° until use.

Western blot analysis

Cells were pelleted and resuspended in SDS sample buffer and run on 15% polyacrylamide SDS gels; Coomassie staining was used to confirm even loading between wells. Gels were equilibrated in transfer buffer (48 mM Tris HCl, 39 mM glycine) and proteins were transferred to PVDF membrane at 10V for 30min using a Trans-blot semi-dry (BioRad) apparatus.

Membranes were blocked overnight at 4° in 5% non-fat powdered milk in TBS-T (20 mM Tris HCl pH 7.5, 137 mM NaCl, 0.05% Tween-20). Blots were washed in TBS-T and incubated at room temperature for 1hr with rocking in rabbit serum diluted 1/10,000 in TBS-T (1% blocking agent). Blots were washed thoroughly and probed with alkaline phosphatase-linked anti-rabbit antibody diluted 1/10,000 in TBS-T (1% blocking agent) for 1hr at room temperature with rocking, followed by thorough washing. Blots were incubated in ECF reagent (Amersham) for 1 minute. Bands were visualized using a STORM 860 scanner and quantified using Image Quant. Several other proteins in addition to Sxy were recognized by the polyclonal antiserum; these were used as internal standards for the quantification of Sxy because of their highly consistent abundance in all culture conditions and growth phases.

Template preparation for RNase analysis

Plasmid DNA was used as template for *in vitro* preparation of mRNAs. The 51nt long untranslated region (UTR) together with the full *sxy* coding region (654 nt) was PCR amplified from genomic DNA isolated from *H. influenzae* KW20 and RR699 (*sxy-1*). Amplicons were digested and cloned into *Apa*I and *Eco*RI restriction sites of pGEM7 (Sigma), adjacent to the T7 promoter, generating plasmids pGEMsxy and pGEMsxy-1 in host strain JM109. pGEMsxy-7 was constructed by PCR amplifying the UTR and coding sequence up to position +272 from the RR854 (sxy-7) chromosome, followed by *Apa*I and *Eco*RI digesting and cloning into pGEM7.

RNA preparation

Wildtype *sxy*, *sxy-1* and *sxy-7* RNAs were prepared by transcription *in vitro* (T7 MEGAscript T7 kit, Ambion) from plasmids linearized at position +272, resulting in 340 nt long run-off transcripts. RNAs were purified from the transcription mix, first by a DNase treatment using a DNA-Free Kit (Ambion) and next by a spin column (RNA Easy kit, Qiagen) following the manufacturer's instructions. At this point each RNA sample was quantified by spectrophotometry, and quality and purity were assessed by agarose gel electrophoresis and $A_{260/280}$ ratios. Next, RNAs (~20 pmol) were dephosphorylated in 100µl reactions at 37°C for 2 hours in 1X reaction buffer using 0.5 U of calf intestinal alkaline phosphatase (Roche). RNAs were recovered by phenol-chloroform purification and ethanol precipitation. Dephosphorylated RNAs (~10 pmol) were labeled in 50µl reactions in 1X reaction buffer using 20 U of T4 polynucleotide kinase (BioLabs, New England) and at least 20 pmol of γ-P^{32} ATP (6000

Ci/mmol, 250 mCi, GE Amersham) at 37°C for 1 hour. Finally, RNAs were purified by a spin column (RNA Easy kit, Qiagen) and eluted in nuclease-free water.

RNA secondary structure mapping

End-labeled RNAs were denatured for 5 minutes at 95°C, allowed to refold for 15 minutes at 37°C, and partially digested with RNase A (0.005 U/ml) or RNase TI (0.05 U/ml) (both from Ambion), and the resulting fragments were resolved on sequencing gels. Both partially digested RNAs and control RNAs (ladders) were prepared following the manufacturers directions. Alkaline digested end-labeled RNA was used as a ladder to help in assigning the bands in the gels to a specific residue in the RNA sequence.

After electrophoresis for 3 hours at 900V and 12mA, gels were dried, exposed to PhosphorScreen overnight and visualized using on a PhosphorImager (Molecular Dynamics). ImageQuant software was used to quantify cleavage intensities at each residue position. Positions +27 and +29 were used as standards to normalize cleavage intensities at all other positions because they were consistently strongly cleaved in independent reactions. To calculate fold differences in cleavage intensities between mutant and wildtype RNA, *sxy*-1 or *sxy*-7 values were divided by wildtype values at each position. At positions where mutant RNAs were more weakly cut than wildtype, wildtype values were instead divided by mutant, and then expressed as a negative value.

***in silico* RNA secondary structure predictions**

Mfold (11), was used to predict 2° structure of the full-length *in vitro sxy* transcript (pGEM7 sequence (15 nt), UTR region (51 nt), and partial coding region (274 nt)); default parameters were used.

Construction of β-galactosidase fusions and enzyme assays

To fuse *lacZ* at *sxy* codon 89, *lacZkan* cassettes (~4.5 kb) were first excised from pLZK80 and pLZK81 with *Bam*HI, then were ligated to *Bcl*I digested pDJM90, generating pLBSF1 and pLBSF2. To eliminate Stems 1 and 3, PCR was used to engineer an *Eco*RI site at position +31 (codon 11); PCR primers: PR6 5'-GAATTCTGTGATTATATCTGTATTGATG, PR15 5'-AGGGAATTCCGCTATCTATATGCTCATCC. The amplicon was digested with EcoRI, ligated to *lacZkan* cassettes, and cloned into *Sca*I + *Bcl*I digested pLBSF1. All gene fusions

were transferred to the KW20 genome by excision from plasmids with *ApaI* + *BamHI* followed by transformation into competent cells.

H. influenzae was grown in sBHI and sampled in duplicate at regular time intervals. For cells in mid to late logarithmic growth ($OD_{600} \geq 0.05$), 0.1 ml of cells was usually sampled; for cells in early exponential growth, larger samples were taken and concentrated by centrifugation. After sampling, cells were immediately pelleted by centrifugation, supernatants were removed, and cell pellets were frozen at -80°C for later assays of β-galactosidase activity. Simultaneously, the main cell culture was assayed for OD_{600} and, in some cases, for cfu/ml.

Quantitative PCR measurement of sxy transcript

Total RNA was isolated from cultures using RNeasy Mini Kits (Qiagen), then was checked for purity and quality by electrophoresis on 1% agarose (1xTAE). RNA was treated twice with a DNA Free kit (Ambion), followed by cDNA synthesis using the iScript cDNA synthesis kit (BioRad). For each PCR primer set, reactions were carried out in duplicate on a 7000 Sequence Detection System (Applied Biosystems) using iTaq SYBR Green Supermix with ROX (BioRad). PCR primers: sxyRTF 5'-TGAACCTTTTACAACGAATGAAT; sxyRTR 5'-ACACAATCTATTACTACGTAAAATCTGATCAG; murGRTF 5'-TGCTTGGGCTGATGTGGTTA; murGRTR 5'-TCCCACTGCTGCAATTTCAC. *murG* RNA served as an internal control for each sample because this gene's expression is constant in the culture conditions used in this study (2). Standard curves were generated using five serial tenfold dilutions of MAP7 chromosomal DNA.

RESULTS

Isolation and characterization of additional hypercompetence mutations in sxy

The original *sxy-1* hypercompetent mutant was isolated from a pool of EMS-mutagenized *H. influenzae* cells created in a search for genes that regulate competence development (3). The present study began by screening more of these cells for mutants that were, like *sxy-1*, transformable in exponential growth, a stage that normally prevents expression of competence genes. This search yielded 4 additional strains with mutations that mapped to *sxy*; the alleles were named *sxy-2*, *sxy-3*, *sxy-4*, and *sxy-5*. As shown in Figure 5.1, all four mutants demonstrated the same 50-fold to 500-fold increased transformation frequencies as the *sxy-1*

mutant, during both exponential growth (OD_{600} 0.2) and late log growth (OD_{600} 1.0). All mutants grew normally in rich medium (sBHI). In MIV starvation medium, mutants and wildtype cells survived equally well and transformed at equally high frequencies.

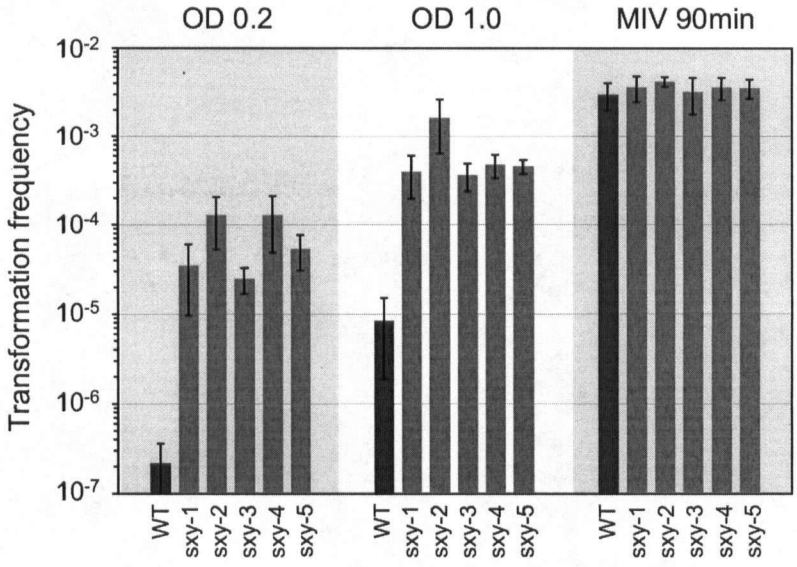


Figure 5.1 Transformation frequencies.
KW20 (blue) and *sxy-1-5* mutants (red) under non-inducing conditions (sBHI at OD_{600} 0.2), moderate inducing conditions (sBHI at OD_{600}) and strong inducing conditions (1.0 and MIV).

Sequencing revealed that each strain carried a distinct single point mutation in *sxy*; these are shown in Figure 5.2. The *sxy-2* mutation ($G_{51}A$) is a silent substitution in the coding region, only 4bp upstream of the *sxy-1* mutation ($G_{55}A$, $V_{19}I$). The other three mutations are clustered outside the coding region, near the 5' end of the 51 nt untranslated region (UTR) (*sxy-3*, $C_{-38}T$; *sxy-4*, $T_{-37}C$; *sxy-5*, $G_{-36}A$).

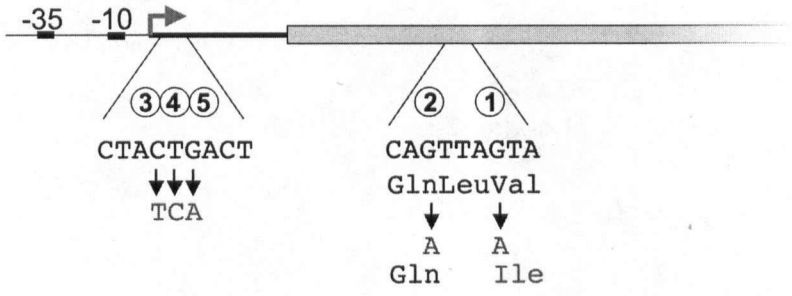


Figure 5.2 Locations of key features and mutations in the *sxy* gene.
Transcriptional controls (-10, -35) are shown relative to the transcription start site (13). Sequences and circled numbers identify *sxy* hypercompetence mutations.

Because these mutations did not alter the Sxy protein sequence, site-directed mutagenesis was used to confirm that no other mutations, either in *sxy* or elsewhere in the genome, were responsible for the hypercompetence phenotypes. As had been done for *sxy-1* (4), each of the four mutations was re-created in a *H. influenzae sxy* plasmid cloned in *E. coli*, and introduced into a wildtype *H. influenzae* chromosome by transformation; these mutants all had phenotypes identical to the originals and were used in the experiments described here. This confirmed that all of the four new hypercompetence mutations increased competence without changing the sequence of Sxy or any other protein. We thus hypothesized that all five mutations acted by altering control of *sxy* expression rather than by changing Sxy function. As Sxy is an activator of competence genes, we predicted that the mutations would cause hypercompetence by increasing rather than decreasing *sxy* expression.

Hypercompetence mutations lead to elevated Sxy under non-inducing and semi-inducing conditions

To compare Sxy abundance between wildtype and mutant cells, we generated polyclonal anti-Sxy antibodies and used western blot analysis to quantify protein levels. In exponential growth (OD_{600} 0.2) all mutants had elevated Sxy levels, with 7-16 fold more protein than wildtype cells (Fig. 5.3A; light green bars). In late log phase (OD_{600} 1.0) the difference was even more striking, with mutants having 13-25 fold more Sxy protein than wildtype cells (Fig. 5.3A, dark green bars). Figure 5.3B graphs transformation frequencies as a function of Sxy protein levels for wildtype and mutant cells in log and late-log growth. The strong positive correlation between Sxy abundance and transformation frequencies suggests that Sxy levels limit competence development during growth in rich medium, and that changes in the amount of Sxy are responsible for the hypercompetence of the *sxy* mutants. The direct correlation also suggests that Sxy activity is not affected by allosteric regulation or post-translational modification.

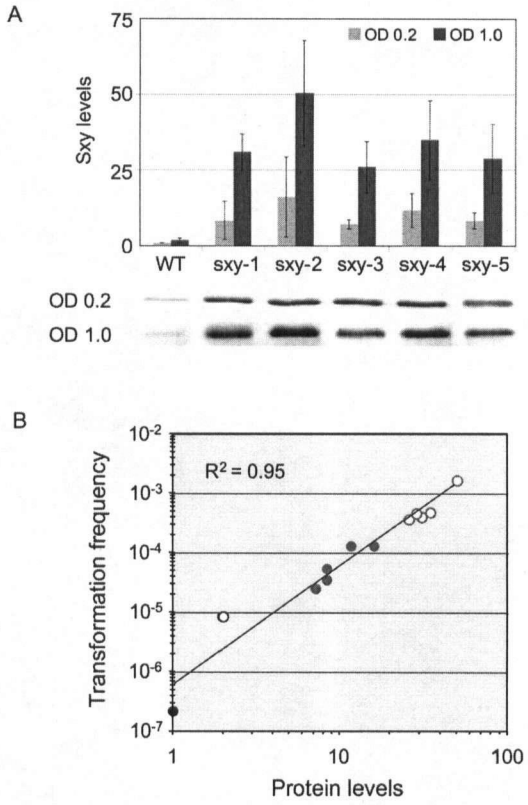


Figure 5.3 Analysis of Sxy levels in wildtype and mutant cells under different growth conditions.

A. Quantitation of Sxy in wildtype and hypercompetent mutants in log (OD_{600} 0.2) and late log (OD_{600} 1.0) growth; values expressed relative to wiltype cells in log growth. The average and standard deviation of four independent cultures are shown in the graph. The bands below the graph show Sxy protein detected by Western blotting. **B.** Transformation frequencies as a function of Sxy protein levels for wildtype cells (blue) and *sxy* hypercompetent mutants (red), in sBHI at OD_{600} 0.2 (solid circles) and OD_{600} 1.0 (open circles).

How do the *sxy-1-5* mutations cause increased Sxy production? Their locations rule out several possible modes of action. The mutations do not improve the affinity of the core promoter elements (-10 and -35 sequences), nor create a more efficient start codon or Shine-Dalgarno sequence, so they are unlikely to act by changing factors that determine baseline expression. Further, the mutations are unlikely to act by changing the binding site for a transcription factor, as they are outside the promoter and spread over 94 bp of transcript sequence.

The clustering of the mutations into two regions suggested that mRNA secondary structure might play a role in regulation. Examination of *sxy* mRNA for possible base pairing between these regions revealed a long stretch of potential base pairing between positions -43 to -25 of the UTR and positions +42 to +60 of the coding region, with only a single 2bp bubble, as shown in

Figure 5.4. All 5 hypercompetence mutations fall within this predicted stem. Moreover, each of the mutations eliminates a base pair within this stem, so that each is expected to destabilize the secondary structure. Analysis of this region with the RNA-folding program Mfold supported this folding model, and also predicted pairing between segments internal to this stem, creating two additional stems and three loops, as shown in Fig. 5.5A.

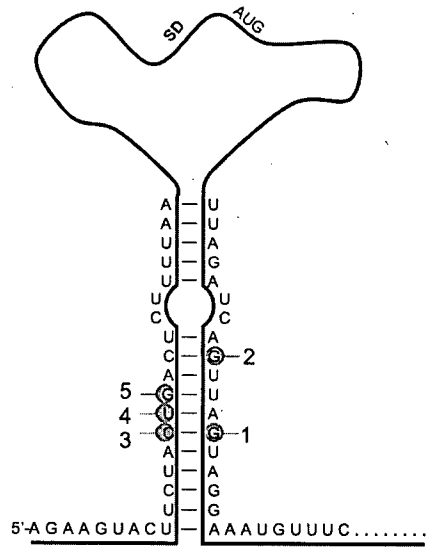


Figure 5.4 Proposed *sxy* secondary structure.
The numbered circles show the locations of the hypercompetence mutations. SD: Shine-Dalgarno site; AUG: start of translation.

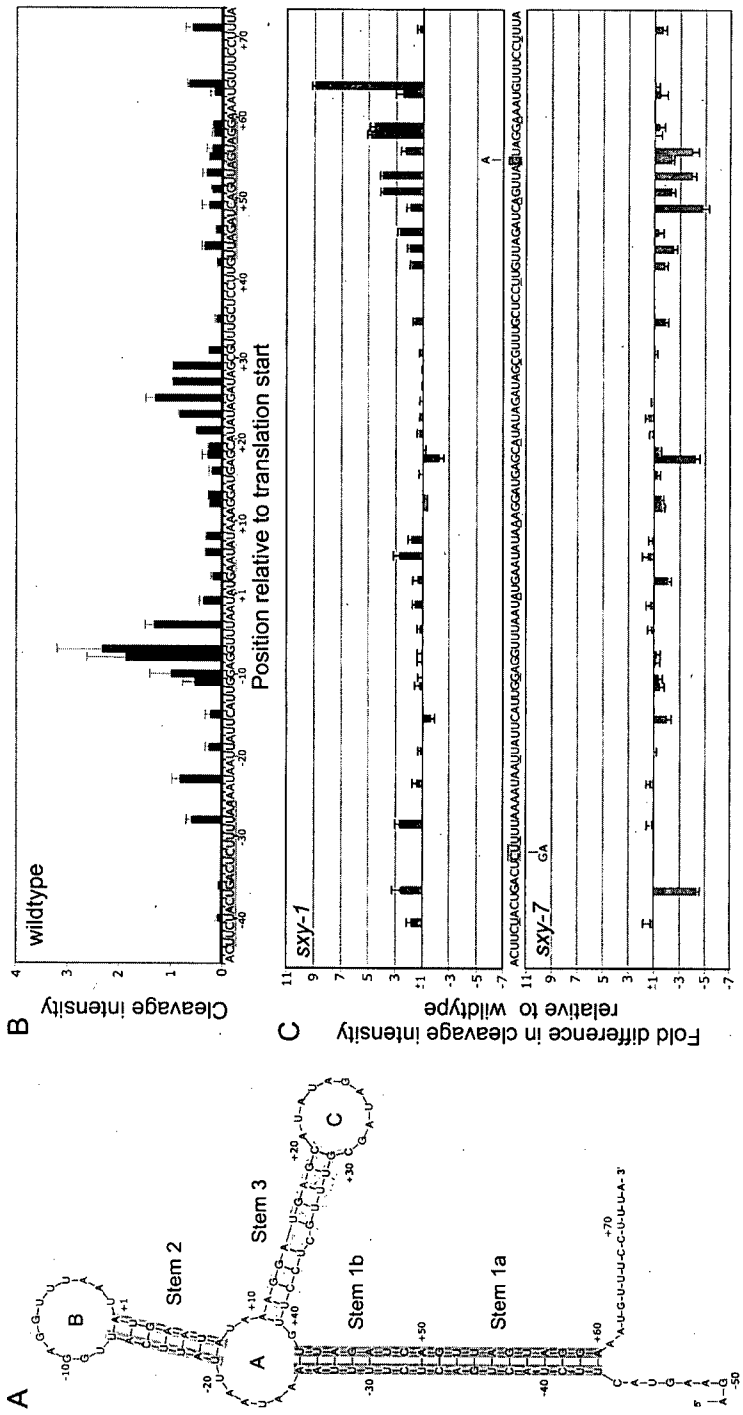


Figure 5.5 RNase analysis of *sxy* mRNA secondary structure.
A. Secondary structure predicted by Mfold. **B.** Cleavage intensity of *sxy* mRNA by single strand specific nucleases RNase A and RNase T1. Colours correspond to stem regions shown in A. **C.** Fold difference in cleavage intensity of mutant RNAs relative to wildtype.

Nuclease mapping confirms the predicted *sxy* mRNA 2° structure

Nuclease mapping was used to test whether *sxy* mRNA folds into the predicted 2° structure *in vitro*, and to test whether hypercompetence mutations alter RNA folding. We first examined cleavage of wildtype *sxy* RNA by the structure-specific ribonucleases RNase T1 and RNase A, using a cloned *sxy* fragment extending from base -51 to base +272. RNase T1 cuts specifically at single stranded Gs, while RNase A cuts single stranded Cs and Us. Fig. 5.5B shows the cleavage intensities of all scorable positions between positions -51 and +71, normalized to positions +27 and +29, which were consistently cleaved. (Data for some Cs and Us are not shown because they were not cut by RNase A even when the RNA was denatured.) The strong cleavages between positions -10 and -4, and between positions +21 and +29, confirm that loops B and C form *in vitro*, and that they are separated by segments that are protected by pairing. The moderate cleavage at position -23 is consistent with the presence of loop A. Only three positions in the upstream (proximal) side of Stem 1 are informative; positions -36 and -40 are protected but position -28 is moderately cleaved, suggesting that Stem 1b may be weak. The segment that would form the distal side of Stem 1 (+43-+71) has more informative positions; these are consistently protected except for positions +64 and +71, which are moderately cleaved.

The nuclease-assay support for Stems 2 and 3 and Loops B and C suggests that the *sxy* Shine-Dalgarno site and start codon may by sequestered within a small loop and stem respectively, likely preventing the initiation of translation. The biochemical evidence for Stem 1 is fairly strong, with most of the informative positions protected from cleavage. Importantly, of the sites of the five hypercompetence mutations, the three that are scorable in these assays are all strongly protected, supporting the hypothesis that they normally are paired.

We then examined mutant *sxy-1* RNA; Figure 5.5C (red bars) shows the effect of the mutation. (Note that this figure shows ratios of cleavages of *sxy-1* and *sxy*+ RNAs). The expected destabilization of Stem 1b by the loss of the base pair between positions -38 and +55 is confirmed by the increased cleavage of positions -41 and -36 and positions +50 to +60. Modest increases in nuclease sensitivity were also seen in Stem 1b (position -28) and Stem 2 (positions +6 and +8). Position +64 was very strongly cleaved.

Mutations that strengthen Stem I reduce translation

The definitive test of whether a mutant phenotype results from disruption of base pairing is creation of compensatory mutations that restore the hypothesized base pairing. The test is especially clear here, as the *sxy-1* and *sxy-3* mutations make complementary substitutions disrupting the same proposed base pair. If both do increase *sxy* expression by destabilizing the secondary structure, then a double mutant carrying both substitutions will have base pairing restored and thus will have a more normal phenotype (lower competence) than either single mutant, rather than the more extreme phenotype expected if the mutations increase expression in some other way. The desired double mutant, *sxy-6*, was created by site-directed mutagenesis in *E. coli*, followed by transformation into the *H. influenzae* chromosome. This combined the *sxy-1* and *sxy-3* mutations to generate an A:U pair where wildtype *sxy* has a G:C pair. Figure 5.6 shows that *sxy-6* cells produced wildtype levels of Sxy protein, much less than either parent mutant, confirming that the *sxy-1* and *sxy-3* mutations act by disrupting base pairing. Consistent with this, transformation assays showed that the *sxy-6* mutant has a transformation frequency below wildtype (not shown).

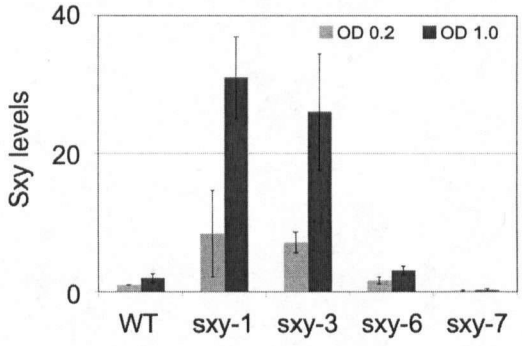


Figure 5.6 Sxy protein in wildtype and in *sxy-1*, *sxy-3*, *sxy-6*, and *sxy-7* mutants.
Quantitation of Sxy protein levels in wildtype and mutants in log (OD_{600} 0.2) and late log (OD_{600} 1.0) growth; values expressed relative to wiltype cells in log growth. The average and standard deviation of four independent cultures are shown in the graph. The values for wildetype, *sxy-1* and *sxy-3* are reproduced from Figure 5.3 to facilitate comparison.

To further characterize the ability of base pairing to limit *sxy* expression, a second mutant with enhanced base pairing was constructed. In *sxy-7*, two adjacent substitutions ($C_{-32}G$ and $U_{-31}A$) create two new base pairings at the site of the 2bp bubble separating Stems 1A and 1B, so Stem 1 has 18 contiguous base pairings. Figure 5.5C (green bars) show that this change strongly reduced RNase cleavages at positions -36 and +51-+56 (Stem 1B), -17 and +3 (Stem 2), +18

and +35 (Stem 3) of the RNA (again the values are relative to those in Fig. 5.5B); the generally decreased cleavage in the entire region indicates stronger base pairing throughout. As predicted, C_{+49}, the predicted pairing partner of G_{-32} was not cleaved. Sxy protein was barely detectable in this mutant (Fig. 5.6) and cells could not be transformed even after transfer to MIV (not shown). Together the *sxy-6* and *sxy-7* mutations confirm that base pairing in Stem I limits *sxy* expression and competence development.

How does mRNA 2° structure regulate *sxy* expression?

In principle, the secondary structure of *sxy* mRNA could limit production of Sxy protein by interfering with elongation of transcription or by reducing the resulting mRNA's stability or translatability. Results of two independent methods of investigation (measurements of β-galactosidase production from *sxy::lacZ* fusions and direct measurements of *sxy* RNA and protein levels) agree that the structure affects both accumulation and translation efficiency of *sxy* mRNA.

The relative impacts of the *sxy* secondary structure on transcription and translation were investigated using transcriptional and translational fusions to the *E. coli lacZ* gene, constructed for wildtype *sxy* and for a truncated *sxy* lacking sequence needed for formation of Stems 1 and 3 (diagramed in Figure 5.7). Fusions 1 and 2 join *sxy* to *lacZ* at *sxy* codon 89 (nucleotide +265), maintaining all of the secondary structure shown in Figure 5.5A. Fusions 3 and 4 join *sxy* to *lacZ* at *sxy* codon 11 (nucleotide +31 of Figure 5.5A), eliminating the distal strands of Stems 1 and 3.

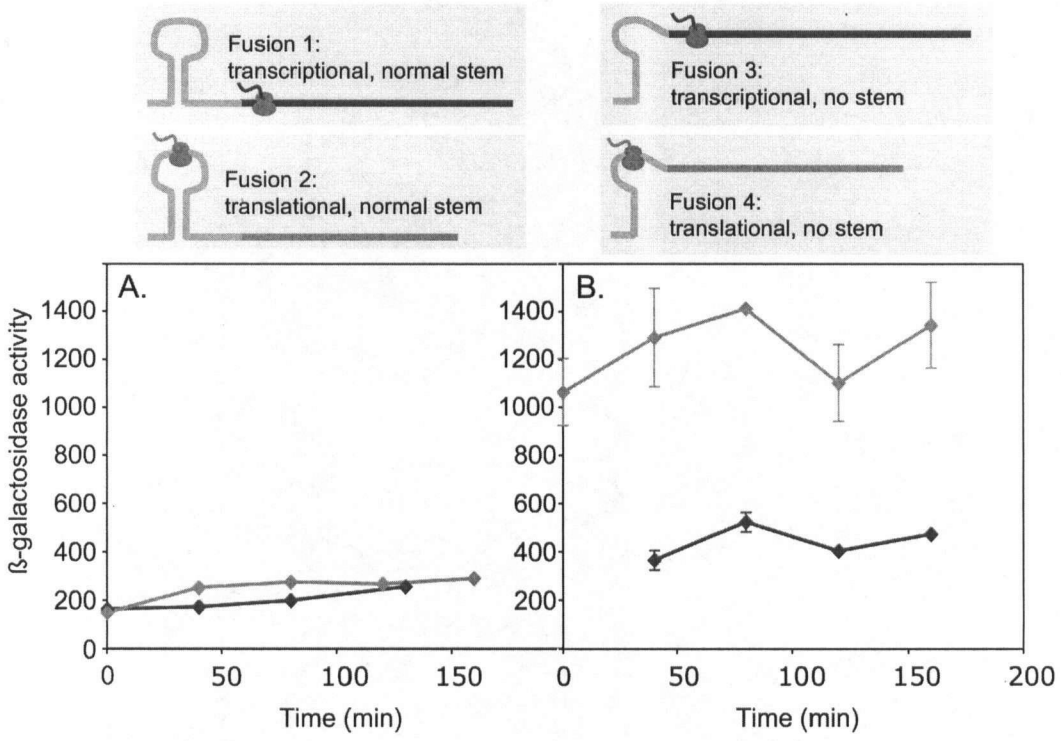


Figure 5.7 Expression from *sxy*::*lacZ* transcriptional and translational fusions.
A. fusions to sxy codon 89; **B.** fusions to *sxy* codon 11. Purple, transcriptional fusions; green, translational fusions. Each point is the mean of two replicate cultures. Error bars representing the range of the replicates are shown only where the range was ≥100 Miller units. The OD_{600} of each sBHI culture increased from 0.085 to ~1.2 over the course of the experiments.

Expression from fusion 1 (purple points) revealed that transcription from the *sxy* promoter is quite stable during exponential growth and early stationary phase in rich medium (Fig. 5.7A). The absence of the *lacZ* translation start site in fusion 2 (Fig. 5.7A, green points) did not significantly change the amount of β-galactosidase activity produced, indicating that the *sxy* and *lac* translation start sites have comparable activities. Figure 5.7B shows that elimination of Stems 1 and 3 increased expression from the transcriptional fusion two-fold (purple points). The 5-fold increase in expression from the translational fusion (green points) therefore represents a 2.5-fold increase in translatability.

A parallel experiment directly tested whether the increased Sxy protein in hypercompetence mutants results from changes in accumulation and/or translatability of *sxy* mRNA. Figure 5.8 plots protein abundance (data from Fig. 5.3A) as a function of mRNA abundance measured by quantitative PCR. When wildtype cells transition from log to late-log growth, the amount of *sxy* transcript doubles, and this results in a doubling of protein abundance; the lower dashed line shows this predicted relationship. Hypercompetence mutations cause a moderate increase in *sxy*

mRNA relative to wildtype in both log and late-log growth. If this was the only effect of the mutations, their points should fall further along the lower dashed line. Instead the points fall well above this line, indicating that more protein is produced from each mRNA.

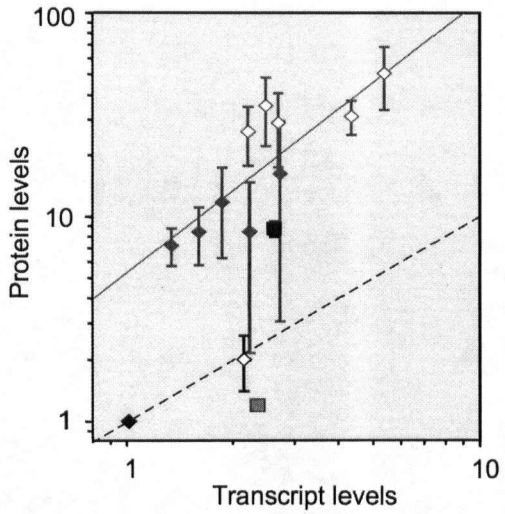


Figure 5.8 Effect of *sxy* mutations on mRNA and protein levels.
Sxy protein levels as a function of transcript levels in wildtype cells (blue) and hypercompetent mutants (red). Solid diamonds, cells in sBHI at OD_{600} 0.2; open diamonds, cells in sBHI at OD_{600} 1.0; solid square, cells in MIV; solid green square, wildtype cells in MIV+1 mM AMP. All values are plotted on log scales and are expressed relative to protein and transcript levels in wildtype OD_{600} 0.2. The dashed line indicates the linear relationship between protein and transcript levels expected if the mutations do not alter translatability of *sxy* mRNA, and the red line is the best fit to the mutant data points.

The data points for hypercompetent mutants, like those for KW20, are from cells growing in rich medium. The line of best-fit through hypercompetent data points is almost parallel to that for KW20, indicating a linear relationship between transcript and protein levels. However, the position of the line implies that the mutant mRNAs are translated on average 5-fold more efficiently. This suggests that the reason MIV medium induces maximum competence in wildtype cells might be that it releases the translation limitation caused by mRNA 2° structure in rich medium. The large blue square in Figure 5.8 shows that wildtype cells have disproportionately elevated levels of Sxy protein relative to *sxy* transcript after 90 minutes in MIV, consistent with relaxation of translational controls.

The expression of two competence genes, *comA* and *rec2*, is inhibited by purine nucleotides (12). We tested whether this repression arises because of changes in *sxy* transcription or translation efficiency when nucleotide pools are high. The green square in Figure 5.8 shows the effect of adding 1 mM AMP to MIV starvation medium. After 90 minutes, *sxy* mRNA levels

are almost as high as in plain MIV (compare green and blue squares in Figure 5.8). However, very little Sxy protein is produced after 90 minutes, suggesting that translation of Sxy is repressed when AMP pools are high.

CRP and cAMP strongly induce *sxy* transcription

We have previously observed a burst of *sxy* transcription immediately upon transfer of wildtype cells from log phase growth to MIV (2), but the effector of this regulation has not been identified. The *sxy* promoter was originally annotated as having two CRP-binding sites (13), and we wished to determine whether CRP stimulates this burst of transcription in MIV. First we scored the putative CRP sites for goodness-of-fit with 58 experimentally determined *H. influenzae* CRP sites, as previously described (1). One scored as an excellent CRP site and is positioned such that it may activate *sxy* transcription (data not shown), however the apparent second site arose by a sequencing error and the authentic sequence is not different from background.

To test whether CRP induces *sxy* in MIV, we measured *sxy* transcript in a *cyaA* mutant that cannot synthesize CRP's allosteric effector cAMP. Transcription was induced only slightly in *cyaA*$^-$ mutants (Fig. 5.9, grey line). Adding 1mM cAMP resulted in very strong induction of *sxy* (blue line), indicating that CRP does stimulate the *sxy* promoter. The promoter could still be induced by cAMP added after 20 minutes in MIV (light blue line), but the amount of *sxy* transcript still fell back to pre-induction levels after 40 minutes in MIV. Thus, CRP is a strong inducer of *sxy* expression, but is overridden by a repressing mechanism even when cAMP levels remain high. This mechanism is unlikely to be auto-repression of the *crp* gene by activated CRP, because *sxy* repression does not depend on when cAMP was added.

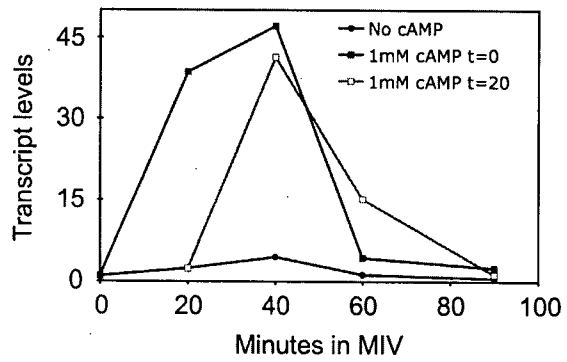


Figure 5.9 Control of *sxy* transcription by cAMP-CRP.
Cells lacking adenylate cyclase were cultured in MIV ± cAMP; *sxy* transcript was measured using real-time PCR.

DISCUSSION

We have identified and characterized an extensive 5' stem-loop structure in the *sxy* transcript that negatively regulates this competence-inducing transcription factor. Mutations that destabilize the 2° structure lead to a moderate increase in *sxy* mRNA level and more efficient translation, causing greatly elevated competence under otherwise non-inducing conditions. These large phenotypic effects arise from minor perturbations in mRNA 2° structure; removing a base pair in Stem 1 (*sxy-1*) destabilizes the structure while adding two base pairs (*sxy-7*) stabilizes folding. *In vivo*, the 2° structure is likely to be dynamic, subject to interplay between the rate of transcription and ribosome loading.

The increase of Sxy protein relative to *sxy* mRNA in hypercompetent mutants (Fig. 5.8) is most readily explained by a simple model in which the wildtype mRNA 2° structure limits initiation of translation. Extensive regions of double-stranded RNA at or near the Shine-Dalgarno (SD) site and start codon, such as the 2° structure we have detected in *sxy* mRNA, are known to preclude ribosome binding and consequently translation. Ribosome binding requires that a 35-50nt segment including the SD site be free of stable 2° structure (14). Dynamic modeling of *sxy* mRNA folding using the RNA Kinetics server (http://www.ig-msk.ru/RNA/kinetics/) predicted that the segment of *sxy* mRNA containing the SD site and the start codon remains largely unstructured until more than 100nt have been synthesized, thus providing a suitable landing platform for ribosomes. Our structural analysis of *sxy-1* RNA demonstrated a significant reduction in base pairing, likely to facilitate binding of ribosomes to longer transcripts. The observed moderate increase in transcript abundance in hypercompetent mutants would, under this hypothesis, be due to greater occupancy and protection of *sxy-1* mRNA by ribosomes rather than increased promoter activity.

The increased ratio of *sxy* protein to mRNA when cells are incubated in MIV starvation medium suggests that the 2° structure does more than establish a baseline level of translation. Rather, it appears to play a sensory role whereby efficient translation is made conditional on nucleotide starvation. In a similar fashion, *E. coli*'s pyrimidine biosynthetic genes *pyrB* and *pyrI* are translated only when nucleotide depletion causes RNA polymerase to stall and prevent mRNA folding (15). Because nucleotide pools are not limiting in favourable growth conditions, transcription can progress unimpeded, allowing *sxy* to fold before ribosomes load. Once folding of the 2° structure is complete, translation can only initiate during rare unfolding events;

unfolding will be more frequent if mutations are present that weaken the structure. Under this model, transcriptional controls on *sxy* expression are released when cAMP levels rise as cells approach stationary phase in rich medium, but translational controls are only effectively released when nucleotide pools are depleted by transfer to MIV.

Other roles for the 2° structure are also possible. Base pairing in the 5' end of nascent *sxy* transcripts could cause transcription to pause or stall, and such attenuation mechanisms can be responsive to regulatory signals such as the availability of nucleotides or amino acids. A role for RNase E can probably be ruled out because weakening the 2° structure leads to elevated transcript levels, contrary to expectations if RNaseE targeted the *sxy* transcript. The 2° structure is unlikely to act as a purine or other riboswitch, as it has no similarities to the well-conserved structures of known riboswitches (16).

Wildtype cells in MIV produce less Sxy protein but become more competent than hypercompetent mutants in rich medium. This suggests that Sxy is not the only factor limiting competence in MIV, and that one or more additional signals induce competence and/or relieve competence repression. MIV-treated cells produce more cAMP than cells in rich medium (17), and this likely causes elevated CRP activity at the CRP-S promoters of competence genes. In addition, PurR repression of purine biosynthesis genes is relaxed in MIV (2); at least one essential competence gene has a candidate PurR binding site in its promoter and we are currently testing whether PurR represses competence when nucleotide pools are high. Understanding the interplay of signals transduced by CRP/cAMP, Sxy, and possibly PurR will clarify how nutritional sensing controls DNA uptake.

REFERENCES

1. Cameron,A.D. and Redfield,R.J. (2006) Non-canonical CRP sites control competence regulons in Escherichia coli and many other gamma-proteobacteria. *Nucleic Acids Res.*, **34**, 6001-6014.
2. 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.
3. Redfield,R.J. (1991) sxy-1, a Haemophilus influenzae mutation causing greatly enhanced spontaneous competence. *J. Bacteriol.*, **173**, 5612-5618.
4. 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.
5. 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.
6. Postma,P.W., Lengeler,J.W. and Jacobson,G.R. (1996) Phosphoenolpyruvate:carbohydrate phosphotransferase system. In Neidhardt,F.N., et al. (ed.), *Escherichia coli and Salmonella typhimurium,* Washington, D.C, Vol. II, pp. 1149-1174.
7. Macfadyen,L.P., Ma,C. and Redfield,R.J. (1998) A 3',5' cyclic AMP (cAMP) phosphodiesterase modulates cAMP levels and optimizes competence in Haemophilus influenzae Rd. *J. Bacteriol.*, **180**, 4401-4405.
8. Poje,G. and Redfield,R.J. (2003) Transformation of Haemophilus influenzae. *Methods Mol Med,* **71**, 57-70.
9. Poje,G. and Redfield,R.J. (2003) General methods for culturing Haemophilus influenzae. *Methods Mol Med,* **71**, 51-56.
10. Ausubel,F.M. (1995) Current protocols in molecular biology. J. Wiley & Sons, Inc, Brooklyn, NY.
11. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406-3415.
12. 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.
13. Zulty,J.J. and Barcak,G.J. (1995) Identification of a DNA transformation gene required for com101A+ expression and supertransformer phenotype in Haemophilus influenzae. *Proc Natl Acad Sci.U S A,* **92**, 3616-3620.
14. de Smit,M.H. and van Duin,J. (2003) Translational standby sites: how ribosomes may deal with the rapid folding kinetics of mRNA. *J. Mol. Biol.*, **331**, 737-743.
15. Donahue,J.P. and Turnbough,C.L.J. (1994) Nucleotide-specific transcriptional pausing in the pyrBI leader region of Escherichia coli K-12. *J. Biol. Chem.*, **269**, 18185-18191.

16. Mandal,M. and Breaker,R.R. (2004) Gene regulation by riboswitches. *Nat Rev Mol Cell Biol,* **5**, 451-463.

17. Macfadyen,L.P. (1999) PTS regulation of competence in Haemophilus influenzae. PhD Thesis, University of British Columbia.

18. Alexander,H.E. and Leidy,G. (1951) Determination of inherited traits of H. influenzae by desoxyribonucleic acid fractions isolated from type-specific cells. *J. Exp. Med.,* **93**, 345-359.

19. Barcak,G.J., Chandler,M.S., Redfield,R.J. and Tomb,J.F. (1991) Genetic systems in Haemophilus influenzae. *Methods Enzymol.,* **204**, 321-342.

CHAPTER SIX

General discussion

This thesis is about the molecular mechanisms that *H. influenzae* uses to link environmental and physiological signals to responses at the gene level. Along with my collaborators, I have demonstrated that competence genes belong to a CRP-dependent regulon that is distinguished by specialized CRP-binding sites called CRP-S sites (Chapter 2). CRP-S sites are unusual because they exhibit a seemingly counter-intuitive approach to promoter activation. Although CRP-S are low affinity sites for their cognate transcription factor (CRP), these sites effectively integrate two signals in a single on/off switch by requiring the concerted activity of both CRP and Sxy for maximal promoter stimulation (Chapter 4). Expression of the *sxy* gene was found to be regulated by two signals; one is the sugar-starvation response conveyed by CRP while the other may arise through the influence of nucleotide starvation on the rate at which *sxy* transcript is translated (Chapter 5). These findings are of broad interest because competence genes and their regulatory mechanism are conserved in several γ-proteobacterial families (Chapter 3).

Together, these insights provide a greater understanding of how multiple environmental and physiological signals are integrated to induce natural competence in *H. influenzae*. Finding that CRP acts directly at CRP-S promoters allies competence genes with several hundred other genes belonging to *H. influenzae*'s global CRP regulon (1). In addition, the body of knowledge surrounding CRP function will greatly inform and guide future research into competence gene regulation in *H. influenzae* and other γ-proteobacteria.

EXPANDING THE GLOBAL CRP REGULON

CRP was first identified as a regulator of diauxic growth in *E. coli* (2). The modern era of genome-wide analysis has elevated CRP from these modest beginnings to be recognized as the chief global regulatory protein in *E. coli* (3). The textbook model of CRP function depicts the lone regulator binding DNA and recruiting RNA polymerase (RNAP). This simple model has been elaborated in light of CRP's interactions with CytR, MelR, and other proteins at various *E. coli* promoters. At the *cdd* promoter, CytR can alter CRP site selectivity by displacing CRP from its preferred "CRP2" site to the adjacent "CRP3" site (4). In contrast, MelR assists CRP binding to a low affinity site at the *melAB* promoter (5, 6). Thus, if Sxy is discovered to bind

specific sites in CRP-S promoters, the MelR model may be the most informative for understanding how Sxy improves CRP binding to CRP-S sites.

In Chapter 4 we propose that Sxy interacts with CRP in solution. This contrasts with the *cdd* and *melAB* promoters where CytR and MelR each bring their own signal to promoter DNA; at these modular promoters, it is the ensuing protein-protein interactions with CRP that influence transcription initiation. Sxy binding to CRP in solution would remove promoter DNA as the only focal point of CRP's co-regulators. Instead, CRP would itself be a nucleating centre that integrates multiple signaling pathways before DNA interactions are initiated. More profoundly, Sxy may be a second allosteric effector of CRP.

Discovery of Sxy-targeted CRP-S sites raises the possibility that CRP has additional, as-yet undetected types of binding site that define unconventional regulons. This may explain the recent observation that most CRP precipitated from *E. coli* is bound to regions that lack previously identified CRP sites (7).

Regulon hierarchy

Gene promoters often incorporate regulatory signals conveyed both by proteins with far-reaching influence and by proteins with small, local effects. This architecture of overlapping regulons is usually described as a hierarchy crowned by global regulators (3, 8). For example, CRP is a master transcription factor that conveys a signal of sugar starvation to hundreds of gene promoters; activation of a transcriptional unit within the global CRP-regulon depends on whether the cognate local regulator also receives a signal indicating that gene induction is favourable.

The *lac* operon presents the best-studied example of transcriptional regulation by CRP and a co-regulator. When CRP signals sugar starvation at the *lacZYA* promoter, transcription does not occur unless the LacI repressor senses the presence of the operon's specific substrate, lactose. Sxy, like LacI, is a local regulator whose function is restricted to induction of CRP-S promoters.

Even though sugar utilization genes such as *lacZYA* are the prototypical members of the CRP regulon, competence genes should be considered equally typical members. In fact, competence genes have been maintained as core members of the CRP regulon ever since CRP first emerged as the global regulator of a sugar/energy starvation response in the common ancestor of the *Pasteurellaceae*, *Enterobacteriaceae*, and *Vibrionaceae* (Chapter 3). The *lac* operon, on the

other hand, may have been present in the common ancestor of the *Enterobacteriaceae* but has since been lost in most lineages (9).

Recently, bacterial regulatory networks have been found to demonstrate great flexibility over evolutionary time (10-12). The CRP-S regulon has remained relatively cohesive over many millions of years, probably because it unites genes that contribute to a common task and, in the case of DNA-binding and uptake, it unites genes that encode multi-subunit protein complexes. Understanding the common goal of genes in this regulon will shed light on a process as fundamental to the cell as sugar metabolism.

Regulation of *sxy* by CRP

CRP regulates many transcription factors in *E. coli*, including most of its co-regulators, making it hard to resolve direct from indirect regulation of many genes (3). Consequently, prior to this thesis work it was unknown whether CRP directly activates competence genes or whether it activates Sxy, which in turn stimulates competence. Figure 7 in Chapter 3 shows that artificial induction of *E. coli.sxy* only stimulates competence genes in the presence of *crp*, clearly demonstrating that both proteins are simultaneously required for competence gene expression. Moreover, inducing *H. influenzae* CRP in rich medium does not result in a concomitant induction of *sxy*, therefore Sxy is sensitive to a CRP-independent signal (13).

CRP-DNA interactions and binding site recognition

In vitro, *Hi*CRP only bound to two of the five promoter DNAs tested, P_{mgl} and P_{pli}-*N*, both of which are the only promoters containing the perfect core sequence $T_4G_5T_6G_7A_8$ in both half sites. *In vivo* however, very few *H. influenzae* CRP sites have the perfect $T_4G_5T_6G_7A_8$ sequence (Chapter 2). Thus, *Hi*CRP may depend on protein cofactors for binding at all but the best CRP sites. An intriguing explanation of *Hi*CRP's low affinity for DNA is that as *H. influenzae* adapted to the metabolically stable niche provided by its obligate human host, it lost the ancestral regulatory diversity still found in *E. coli*. In the course of losing CRP-regulated genes and co-regulators from the genome, *Hi*CRP lost many protein-protein and protein-DNA interactions. The absence of functional constraints may have allowed *Hi*CRP to drift towards low DNA affinity; now it can bind stably only to very high quality CRP sites *in vitro*. To determine whether this is a common biological phenomenon, this hypothesis awaits

experimental characterization of transcription factors in bacterial lineages with diverse genome sizes; transcription factors in reduced genomes are predicted to exhibit similar loss-of-function.

FNR does not regulate competence

FNR is the only other member of the CRP family in *H. influenzae*. In *E. coli* the two proteins have very similar binding sites. For example, mutation of both halves of the *lacZYA* CRP site from $T_4\underline{G}_5T_6G_7A_8$ to $T_4\underline{T}_5T_6G_7A_8$ removes P_{lac} from the CRP regulon and places it in the FNR regulon (14, 15). FNR's binding specificity has not been investigated in *H. influenzae*, raising the possibility that it also binds and regulates CRP-S promoters. However, knocking out *fnr* had no effect on competence in sBHI or MIV (C. Ma, personal communication).

DO CRP AND SXY PHYSICALLY INTERACT?

An outstanding issue is whether Sxy and CRP interact. A recent large-scale analysis of protein complexes in *E. coli* identified several interaction partners for both CRP and Sxy (16), but CRP and Sxy were separated by at least two proteins in the resulting interaction network. However, this study was not designed to detect interactions occurring in nucleoprotein complexes. DNA was removed from protein-protein binding assays, and this may explain why the study failed to detect any of CRP's previously characterized interactions.

Chapter 4 presents evidence that Sxy exerts its influence on gene expression by directly modifying CRP's DNA binding characteristics, perhaps by reducing CRP's dissociation from CRP-S sites. A model is presented in which Sxy is able to function at low cellular concentrations, ensuring that CRP-Sxy complexes constitute only a small fraction of the total cellular CRP pool. Alternatively, Sxy may have evolved a low affinity for CRP to prevent excessive sequestration. In this latter model, Sxy-CRP complexes are transient until a strong promoter complex is formed at CRP-S sites. Further attempts to co-purify CRP and Sxy should include CRP-S promoter DNA as well as RNAP; fixation of nucleoprotein complexes *in vivo* using formaldehyde may improve recovery of CRP-S promoter complexes.

REGULATION OF *SXY* EXPRESSION IN *H. INFLUENZAE*

In Chapter 5 we hypothesize that the extensive secondary structure in *sxy* mRNA makes translation contingent on transcriptional stalling, thus linking gene expression to nucleotide

availability. However, there are other mechanisms by which nucleotide pools are known to influence rates of transcription.

Transcription initiation is rate-limited by the availability of the first nucleotide in the transcript (17). For example, translational machinery is a major consumer of ATP, so transcription of ribosomal (r)RNA is directly repressed when ATP pools are depleted in *E. coli* (18, 19). This is achieved in part by a requirement for high concentrations of the transcript-initiating nucleotides ATP and GTP at rRNA promoters (20). Most cellular transcripts start with ATP, so this mechanism may have a global influence on transcription (21). The *H. influenzae sxy* transcript begins with ATP, however this is unlikely to serve as a primary sensor of nucleotide pools because our result show that elevated ATP pools have the opposite effect of decreasing *sxy* transcript levels (Chapter 5).

At several *E. coli* promoters, nucleotide pools also influence the rate at which RNAP escapes the promoter to enter elongation mode. If RNAP encounters homopolymeric runs of nucleotides during initiation, it is prone to slip and reiteratively transcribe the same sequence of RNA (22-24). Reiterative transcription is more likely to occur when nucleotide pools are high, and slower RNAP movement favours promoter escape as pools are depleted (25). The *sxy* transcript does not contain homopolymeric runs at its 5' end, so reiterative transcription is unlikely to occur.

Is *sxy* regulated by attenuation?

Many bacterial biosynthetic genes are regulated by attenuation in which an intrinsic (also called rho-independent) terminator located at the beginning of a transcript presents a barrier to transcription (reviewed in (26)). At the archetypal *E. coli trp* operon, ribosomes stall on nascent transcripts when $tRNA^{Trp}$ is limiting; this prevents mRNA from folding into a terminator hairpin and allows RNA polymerase to fully transcribe genes required for tryptophan biosynthesis (reviewed in (27)).

Intrinsic terminators are usually short, GC rich hairpins able to form before vacating the RNA polymerase. In contrast, the *sxy* mRNA secondary structure identified in Chapter 5 is too large to assemble before the newly synthesized RNA has exited the polymerase. However, both Stem 2 and 3 can each fold immediately after synthesis and may slow or terminate RNA polymerase. Even if Stems 2 and 3 are barriers to RNA polymerase progress, our discovery of

hypercompetence mutations in both sides of Stem 1 reveals that the entire secondary structure is important for preventing *sxy* expression. Given the apparent regulatory role of the extensive *sxy* mRNA secondary structure, Stems 2 and 3 are unlikely to prematurely terminate transcription. *In vitro* transcription assays and northern blots may be used to test for transcripts terminating after synthesis of Stems 2 or 3.

REGULATION OF COMPETENCE IN *E. COLI*

Long-term survival studies have demonstrated a nutritional role for DNA uptake in *E. coli*, consistent with our findings that CRP induces *E. coli* competence genes such as *ppdD* and *yrfD* (*comA*) (28, 29). In an attempt to study type IV pili in *E. coli*, Sauvonnet et al. (30) conducted an exhaustive search for conditions that induce *ppdD* and *yrfD*, two members of the CRP-S regulon. The inability of varied culture media, growth phase, anaerobiosis, and acid stress to detectably stimulate competence genes indicates that *sxy* is silent under many culture conditions (30). Furthermore, a search of the microarray data compiled at NCBI's Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo/) did not identify any conditions that alter *sxy* expression, but *sxy* transcripts are usually more abundant in M9 and Davis minimal media than in LB (A. Cameron, personal observation). Given that a rapid shift from rich to starvation medium strongly induces competence in *H. influenzae*, a similar induction protocol should be attempted by transferring exponentially growing *E. coli* cells from LB to minimal medium.

H. influenzae cannot synthesize pyrimidine nucleotides *de novo*, so it is immediately starved upon transfer to MIV medium. On the other hand, *E. coli* can synthesize all necessary nucleotides using sugars and amino acids (31, 32). Consequently, experiments to test the effects of nucleotide starvation and supplementation on *sxy* expression in *E. coli* may require the use of nucleotide auxotrophs to ensure experimental control over nucleotide pools.

The only phenotype currently linked to competence genes in *E. coli* is the competitive advantage they provide in long-term culture. Competition experiments should be conducted to test whether *sxy*- cells demonstrate the same reduced fitness as cells lacking other competence genes (28, 29). In addition, although CRP-S gene expression is undetectable in LB medium under standard culture conditions, gene expression should be tested several days after the onset of stationary phase when a competitive advantage begins to emerge. Because *E. coli* is not known to produce extracellular nucleases (28), exogenous DNA may present an economical

source of nucleotides when most nutrients are tied up in dead cell matter. Similar studies of competence regulatory signals and long-term survival on DNA substrates will help resolve whether natural competence also plays a nutritional role in other γ-proteobacteria.

REFERENCES

1. 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.
2. Zubay,G., Schwartz,D. and Beckwith,J. (1970) Mechanism of activation of catabolite-sensitive genes: a positive control system. *Proc Natl Acad Sci U S A*, 66, 104-110.
3. Martinez-Antonio,A. and Collado-Vides,J. (2003) Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr Opin Microbiol*, 6, 482-489.
4. Holst,B., Sogaard-Andersen,L., Pedersen,H. and Valentin-Hansen,P. (1992) The cAMP-CRP/CytR nucleoprotein complex in Escherichia coli: two pairs of closely linked binding sites for the cAMP-CRP activator complex are involved in combinatorial regulation of the cdd promoter. *EMBO J.*, 11, 3635-3643.
5. 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.
6. 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.
7. Grainger,D.C., Hurd,D., Harrison,M., Holdstock,J. and Busby,S.J. (2005) Studies of the distribution of Escherichia coli cAMP-receptor protein and RNA polymerase along the E. coli chromosome. *Proc Natl Acad Sci U S A*, 102, 17693-17698.
8. Ma,H.W., Buer,J. and Zeng,A.P. (2004) Hierarchical structure and modules in the Escherichia coli transcriptional regulatory network revealed by a new top-down approach. *BMC Bioinformatics*, 5, 199.
9. Stoebel,D.M. (2005) Lack of evidence for horizontal transfer of the lac operon into Escherichia coli. *Mol. Biol. Evol.*, 22, 683-690.
10. Lozada-Chavez,I., Janga,S.C. and Collado-Vides,J. (2006) Bacterial regulatory networks are extremely flexible in evolution. *Nucleic Acids Res.*, 34, 3434-3445.
11. Babu,M.M., Luscombe,N.M., Aravind,L., Gerstein,M. and Teichmann,S.A. (2004) Structure and evolution of transcriptional regulatory networks. *Curr Opin Struct Biol*, 14, 283-291.
12. Madan Babu,M., Teichmann,S.A. and Aravind,L. (2006) Evolutionary dynamics of prokaryotic transcriptional regulatory networks. *J. Mol. Biol.*, 358, 614-633.
13. Bannister,L.A. (1999) An RNA secondary structure regulates sxy expession and competence development in Haemophilus influenzae. PhD Thesis, University of British Columbia.
14. Spiro,S., Gaston,K.L., Bell,A.I., Roberts,R.E., Busby,S.J. and Guest,J.R. (1990) Interconversion of the DNA-binding specificities of two related transcription regulators, CRP and FNR. *Mol. Microbiol.*, 4, 1831-1838.
15. Zhang,X.P., Gunasekera,A., Ebright,Y.W. and Ebright,R.H. (1991) Derivatives of CAP having no solvent-accessible cysteine residues, or having a unique solvent-accessible

cysteine residue at amino acid 2 of the helix-turn-helix motif. *J Biomol Struct Dyn,* 9, 463-473.

16. Butland,G., Peregrin-Alvarez,J.M., Li,J., Yang,W., Yang,X., Canadien,V., Starostine,A., Richards,D., Beattie,B., Krogan,N., Davey,M., Parkinson,J., Greenblatt,J. and Emili,A. (2005) Interaction network containing conserved and essential protein complexes in Escherichia coli. *Nature,* 433, 531-537.
17. Walker,K.A., Mallik,P., Pratt,T.S. and Osuna,R. (2004) The Escherichia coli Fis promoter is regulated by changes in the levels of its transcription initiation nucleotide CTP. *J. Biol. Chem.,* 279, 50818-50828.
18. Barker,M.M. and Gourse,R.L. (2001) Regulation of rRNA transcription correlates with nucleoside triphosphate sensing. *J. Bacteriol.,* 183, 6315-6323.
19. Schneider,D.A., Gaal,T. and Gourse,R.L. (2002) NTP-sensing by rRNA promoters in Escherichia coli is direct. *Proc Natl Acad Sci U S A,* 99, 8602-8607.
20. Gaal,T., Bartlett,M.S., Ross,W., Turnbough,C.L.J. and Gourse,R.L. (1997) Transcription regulation by initiating NTP concentration: rRNA synthesis in bacteria. *Science,* 278, 2092-2097.
21. McClure,W.R., Cech,C.L. and Johnston,D.E. (1978) A steady state assay for the RNA polymerase initiation reaction. *J. Biol. Chem.,* 253, 8941-8948.
22. Wagner,L.A., Weiss,R.B., Driscoll,R., Dunn,D.S. and Gesteland,R.F. (1990) Transcriptional slippage occurs during elongation at runs of adenine or thymine in Escherichia coli. *Nucleic Acids Res.,* 18, 3529-3535.
23. Uptain,S.M., Kane,C.M. and Chamberlin,M.J. (1997) Basic mechanisms of transcript elongation and its regulation. *Annu. Rev. Biochem.,* 66, 117-172.
24. Qi,F. and Turnbough,C.L.J. (1995) Regulation of codBA operon expression in Escherichia coli by UTP-dependent reiterative transcription and UTP-sensitive transcriptional start site switching. *J. Mol. Biol.,* 254, 552-565.
25. Cheng,Y., Dylla,S.M. and Turnbough,C.L.J. (2001) A long T. A tract in the upp initially transcribed region is required for regulation of upp expression by UTP-dependent reiterative transcription in Escherichia coli. *J. Bacteriol.,* 183, 221-228.
26. Yanofsky,C. (2000) Transcription attenuation: once viewed as a novel regulatory strategy. *J. Bacteriol.,* 182, 1-8.
27. Yanofsky,C. (2004) The different roles of tryptophan transfer RNA in regulating trp operon expression in E. coli versus B. subtilis. *Trends Genet.,* 20, 367-374.
28. Finkel,S.E. and Kolter,R. (2001) DNA as a nutrient: novel role for bacterial competence gene homologs. *J. Bacteriol.,* 183, 6288-6293.
29. 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.
30. 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.

31. Neuhard,J. and Kelln,R.A. (1996) Biosynthesis and conversion of pyrimidines. In Neidhardt,F.N., et al. (ed.), *Escherichia coli and Salmonella typhimurium,* Washington, D.C, Vol. II, pp. 580-599.

32. Zalkin,H. and Nygaard,P. (1996) Biosynthesis of purine nucleotides. In Neidhardt,F.N., et al. (ed.), *Escherichia coli and Salmonella typhimurium,* Washington, D.C, Vol. II, pp. 561-579.

APPENDIX 1

Genes induced greater than 4 fold upon transfer to MIV

CRE genes

comA	HI0439	competence protein A
comB	HI0438	competence protein B
comC	HI0437	competence protein C
comD	HI0436	competence protein D
comE	HI0435	competence protein E
comE1	HI1008	conserved hypothetical protein
comM	HI1117	competence protein
dprA	HI0985	DNA processing chain A
pilA	HI0299	prepilin peptidase dependent protein D
pilB	HI0298	protein transport protein
pilC	HI0297	protein transport protein
pilD	HI0296	type 4 prepilin-like protein specific leader peptidase
radC	HI0952	DNA repair protein
rec2	HI0061	recombination protein
	HI0365	conserved hypothetical protein
	HI0659	predicted coding region
	HI0660	predicted coding region
	HI0938	predicted coding region
	HI0939	predicted coding region
	HI0940	predicted coding region
	HI0941	predicted coding region
	HI1182/3	predicted coding region
	HI1631	predicted coding region

sxy

sxy	HI0601	DNA transformation protein

CRP-regulated genes

afuA	HI0131	afuA protein
afuC	HI0126	ferric ABC transporter, ATP-binding protein
ansB	HI0745	L-asparaginase II
aspA	HI0534	aspartate ammonia-lyase
cdd	HI1350	cytidine deaminase
cspD	HI1434-1	cold shock-like protein
eda	HI0047	4-hydroxy-2-oxoglutarate/2-deydro-3-deoxyphosphogluconate
aldolase		
fbp	HI1645	fructose-1,6-bisphosphatase
frdA	HI0835	fumarate reductase, flavoprotein subunit
frdB	HI0834	fumarate reductase, iron-sulfur protein
frdC	HI0833	fumarate reductase, 15 kDa hydrophobic protein
frdD	HI0832	fumarate reductase, 13 kDa hydrophobic protein
fucA	HI0611	L-fuculose phosphate aldolase

fucI	HI0614	L-fucose isomerase
fucK	HI0613	fuculokinase
fucP	HI0610	L-fucose permease
fucU	HI0612	fucose operon protein
glgA	HI1360	glycogen synthase
glgB	HI1357	1,4-alpha-glucan branching enzyme
glgC	HI1359	glucose-1-phosphate adenylyltransferase
glgP	HI1361	glycogen phosphorylase
glgX	HI1358	glycogen operon protein
gntP	HI1015	gluconate permease
icc	HI0399	lacZ expression regulator
kdgK	HI0049	2-dehydro-3-deoxygluconokinase
lctP	HI1218	L-lactate permease
malQ	HI1356	4-alpha-glucanotransferase
mdh	HI1210	malate dehydrogenase
mglA	HI0823	galactoside ABC transporter, ATP-binding protein
mglB	HI0822	galactose ABC transporter, periplasmic-binding protein
mglC	HI0824	galactoside ABC transporter, permease protein
moaC	HI1675	molybdenum cofactor biosynthesis protein C
moaD	HI1674	molybdopterin converting factor, subunit 1
moaE	HI1673	molybdopterin converting factor, subunit 2
nagA	HI0140	N-acetylglucosamine-6-phosphate deacetylase
nagB	HI0141	glucosamine-6-phosphate isomerase
nhaC	HI1107	Na+/H+ antiporter
oppA	HI1124	oligopeptide ABC transporter, periplasmic-binding protein
pckA	HI0809	phosphoenolpyruvate carboxykinase
potE	HI0590	putrescine-ornithine antiporter
rbsA	HI0502	D-ribose ABC transporter, ATP-binding protein
rbsB	HI0504	D-ribose ABC transporter, periplasmic-binding protein
sdaA	HI0288	L-serine deaminase
sdaC	HI0289	serine transporter
speF	HI0591	ornithine decarboxylase
sucA	HI1662	2-oxoglutarate dehydrogenase E1 component
sucB	HI1661	2-oxoglutarate dehydrogenase E2, dihydrolipoamide succinyltransferase
uraA	HI1227	uracil permease
uspA	HI0815	universal stress protein A
xylA	HI1112	xylose isomerase
xylF	HI1111	D-xylose ABC transporter, periplasmic-binding protein
xylG	HI1110	D-xylose ABC transporter, ATP-binding protein
xylH	HI1109	D-xylose ABC transporter, permease protein
yhxB	HI0740	phosphomannomutase
	HI0035	conserved hypothetical protein
	HI0048	oxidoreductase
	HI0050	conserved hypothetical transmembrane protein
	HI0051	conserved hypothetical transmembrane protein
	HI0052	conserved hypothetical protein
	HI0053	zinc-type alcohol dehydrogenase

	HI0129	predicted coding region
	HI0145	conserved hypothetical protein
	HI0146	conserved hypothetical protein
	HI0147	conserved hypothetical transmembrane protein
	HI0148	conserved hypothetical protein
	HI0398	conserved hypothetical protein
	HI0592	predicted coding region
	HI0608	conserved hypothetical protein
	HI0804	predicted coding region
	HI1014	conserved hypothetical protein
	HI1016	predicted coding region
	HI1028	conserved hypothetical protein
	HI1029	conserved hypothetical transmembrane protein
	HI1030	conserved hypothetical transmembrane protein
	HI1031	conserved hypothetical protein
	HI1108	aminotransferase
	HI1126-1	predicted coding region
	HI1127	predicted coding region
	HI1245	malate oxidoreductase, putative
	HI1315	predicted coding region
	HI1316	predicted coding region

PurR regulon

cvpA	HI1206	colicin V production protein
purC	HI1726	phosphoribosylaminoimidazole-succinocarboxamide synthase
purD	HI0888	phosphoribosylamine--glycine ligase
purE	HI1615	phosphoribosylaminoimidazole carboxylase, catalytic subunit
purF	HI1207	amidophosphoribosyltransferase
purH	HI0887	phosphoribosylaminoimidazolecarboxamide formyltransferase
purK	HI1616	phosphoribosylaminoimidazole carboxylase, ATPase subunit
purL	HI0752	phosphoribosylformylglycinamidine synthase
purM	HI1429	phosphoribosylaminoimidazole synthetase
purN	HI1428	phosphoribosylglycinamide formyltransferase

TrpR regulon

mtr	HI0287	tryptophan-specific transport protein
trpA	HI1432	tryptophan synthase alpha subunit
trpB	HI1431	tryptophan synthase beta subunit
trpC	HI1389-1	indole-3-glycerol phosphate synthase
trpD	HI1389	anthanilate phosphoribosyltransferase
trpE	HI1387	anthranilate synthase component I
trpG	HI1388	anthranilate synthase component II
trpR	HI0830	trp operon repressor
	HI1430	short chain dehydrogenase/reductase

Other

aspC	HI1617	aspartate aminotransferase
cpdB	HI0583	2',3'-cyclic-nucleotide 2'-phosphodiesterase

dcuB	HI0746	anaerobic C4-dicarboxylate membrane transporter protein
fumC	HI1398	fumarate hydratase, class II
glpF	HI0690	glycerol uptake facilitator protein
glpK	HI0691	glycerol kinase
glpQ	HI0689	glycerophosphoryl diester phosphodiesterase
glpT	HI0686	glycerol-3-phosphatase transporter
glpX	HI0667	glpX protein
mopI	HI1370	molybdenum-pterin binding protein
nanA	HI0142	N-acetylneuraminate lyase
rsgA	HI1384	ferritin
rsgA	HI1385	ferritin
yecK	HI0644	cytochrome C-type protein
	HI0092	predicted coding region
	HI0125	conserved hypothetical protein
	HI0206	5'-nucleotidase, putative
	HI0234	predicted coding region
	HI0668	conserved hypothetical protein
	HI0843	predicted coding region
	HI1189	conserved hypothetical protein
	HI1190	6-pyruvoyl tetrahydrobiopterin synthase, putative
	HI1191	conserved hypothetical protein
	HI1525	molybdate-binding periplasmic protein, putative
	HI1546	impA protein, putative
	HI1728	conserved hypothetical protein
	HI1729	conserved hypothetical protein

Genes repressed greater than 4 fold upon transfer to MIV

adhC	HI0185	alcohol dehydrogenase, class III
betT	HI1706	high-affinity choline transport protein
cyaA	HI0604	adenylate cyclase
dnaJ	HI1238	heat shock protein
fabA	HI1325	3-hydroxydecanoyl-(acyl carrier-protein) dehydratase
fabH	HI0157	beta-ketoacyl-ACP synthase III
hindIIM	HI0513	modification methylase
hindIIR	HI0512	Type II restriction endonuclease
infB	HI1284	translation initiation factor 2
menA	HI0509	1,4-dihydroxy-2-naphthoate octaprenyltransferase
nusA	HI1283	N utilization substance protein A
psd	HI0160	phosphatidylserine decarboxylase proenzyme
rnb	HI1733	exoribonuclease II
rpL11	HI0517	ribosomal protein L11
rpL2	HI0780	ribosomal protein L2
rpL22	HI0782	ribosomal protein L22
rpL23	HI0779	ribosomal protein L23
rpL25	HI1630	ribosomal protein L25

rpL3	HI0777	ribosomal protein L3
rpL4	HI0778	ribosomal protein L4
rpS19	HI0781	ribosomal protein S19
rpS3	HI0783	ribosomal protein S3
secY	HI0798	preprotein translocase SecY subunit
tgt	HI0244	tRNA-guanine transglycosylase
	HI0036	ABC transporter, ATP-binding protein
	HI0184	esterase
	HI0230	conserved hypothetical protein
	HI0282	conserved hypothetical protein
	HI0673	conserved hypothetical protein
	HI0862	conserved hypothetical protein
	HI0864	GTP-binding protein
	HI1051	ABC transporter, ATP-binding protein
	HI1078	amino acid ABC transporter, ATP-binding protein
	HI1079	amino acid ABC transporter, permease protein
	HI1154	proton glutamate symport protein, putative
	HI1259	periplasmic serine protease
	HI1265	conserved hypothetical protein
	HI1282	conserved hypothetical protein
	HI1301	carbonic anhydrase, putative
	HI1424	integrase/recombinase, putative
	HI1436-1	conserved hypothetical protein
	HI1618	ABC transporter, ATP-binding protein
	HI1620	predicted coding region
	HI1621	conserved hypothetical protein

APPENDIX 2

Table 1. Motif sites in *Pasteurellaceae* promoters. *A.a.* and *M.h.* genomes are not annotated.

Gene	Sequence	*H.i.* ortholog
	H. influenzae	
HI0439	TTTTGCGATCCGCATCGTAAAA	*comA*
HI1008	TTTTGCGATCGAGATCGCAAAA	*comE1*
HI1117	TTTTGCGATCTAGATCGCAAAA	*comM*
HI0985	TTTTGCGATCTGCATCGCAAAA	*dprA*
HI0299	TTTTGCGATCAGGATCGCAGAA	*pilA*
HI0952	TTTTACGATATGCATCGCAGAT	*radC*
HI0061	TTTTACGATATGGATCGCAAAA	*rec2*
HI0250	TTTTGCGATCATTATCGCATAT	*ssb*
HI0365	ATTTGCGATCTAGATCGCAAAA	
HI0660	TTTTGCGATCTAGATCGAAAGA	
HI0938	CTTTGCGATACAGATCGCAAAA	
HI1182	TTTTGCGATTTAGATCGAAAAA	
HI1631	TTTTGCGATTCAGATCCCAAAC	
	M. succiniciproducens	
MS1974	TTTTACGATCTTCATTCCAAAA	*comA*
MS0826	CGGAACGAAAATAATGGCAAAA	*comE1*
MS2234	TATTGCGATAAAGATCGAAAAA	*comF*
MS1998	TTCTGCGAGCCGGATCTCAAAG	*comM*
MS0041	TTTTTCGAGCCGTATCGTAAAA	*dprA*
MS0364	TTTTGCGATCCTGCTCGAGAAT	*pilA*
MS1940	TTTTGCGATCCGTTTCAAAAAA	*radC*
MS0931	AAAGGCGATATAAATAGCAGAA	*rec2*
MS0585	AATTGCGAGCATTATCGCATAT	*ssb*
MS1916	AATTGGAATCACTATCGCAAAA	HI0365
MS0724	TATTGCGATCCTGATCGTAAAA	HI0938
MS0939	TTTATCGATCTTCACCGCAAAT	HI1182
	A. actinomycetemcomitans	
Not ann.	TTTTGCGATCCGCATCGAAAAT	*comA*
	TTTTGCGACCGGGATCCCAAAA	*comE1*
	TTTTGCGAGGCGGATCCTAAAC	*comF*
	TTCTGCGATCCCGATCGCAAAA	*comM*
	AATTACGATCCGGATCACAAAT	*dprA*
	TTTTGCGATCGGGATCCCATAA	*pilA*
	TTTTGTGATTCAGTTTCCAATA	*rec2*
	ATTTGCGATAATTATCGCATAT	*ssb*
	TTCTTCGATCCTGATCACAAAA	HI0365
	CTTTGCGATCCTGCTCGCAAAA	HI0938
	ATTTGGGATCGCCGTCGCAAAA	HI1182
	P. multocida	
PM1229	TTTTGCGATCCGCATCGGGAAA	*comA*
PM1665	TTTTTCGATCTTCATCTCAAAA	*comE1*
PM1556	TTTTGCGATGCGTGTCGCAAAA	*comF*
PM1510	TTCTGCGATCTAGATCGTAAAA	*comM*
PM1599	TTTTACGATCATCCTCACAACC	*dprA*
PM0084	TTTTGCGATAAAGATCGAAAAA	*pilA*
PM1152	TTTTGCGATCTTATTTCCAGAG	*radC*

PM0862	AAAAGCGTTATAAATAGCAGAA	*rec2*
PM1950	AATTGCGTTCATTATCGCACAT	*ssb*
PM2007	TTTCACGATCGAGATCGCAAAA	HI0365
PM0965	TTTTGCGATCTGCATCTCAAAA	HI0938
	H. somnus	
Haso1896	TTTTGCGATCCTCATCGTAAAA	*comA*
Haso1520	TTTTGCGATCTTGATCGTAAAA	*comE1*
Haso0188	TTTTGTGATTAAGATCGAGAAA	*comF*
Haso2123	TTTTACGATCCGGATCGCAAAA	*comM*
Haso1155	TTTTTCGACATATCTCGCAAAA	*dprA*
Haso1470	TTTTGCGAGTCGGCTCGCAGAA	*pilA*
Haso1690	TTTTACGATCCAGATCGTAAAA	*pilB*
Haso0903	TTTTGCGATCTTGATCGTAAAA	*pilC*
Haso1869	TTTTGCGATTTTGCACGCAAAA	*radC*
Haso1385	TTTTGTGATTTGTATTCCAAGA	*rec2*
Haso0534	ATTTGCGATCCGGATCGCATAA	HI0365
Hsom0256	TTTTGCGATCTGTATCGTAATT	HI0938
Haso1003	TTTTGCGATCTCTCTCGCAAAT	HI1182
	H. ducreyi	
HD0427	TTTTGCGATCTTCATCGAAAAA	*comA*
HD0650	TTTCTCGATCAAAATCGCAAAA	*comE1*
HD0209	TTTTTCGACTTATATCGCAAAA	*comF*
HD1870	TTTTGCGATCACGATCGTGAAA	*comM*
HD1888	TTTTGTGATCTCAATCGAAAAA	*dprA*
HD1123	TTTTGCGATATAGATCGAATAA	*pilA*
HD0732	TTTTGCGATCTCCCTCGAAAAA	*radC*
HD1256	TTTTGCGATCTTGATCGAAATT	*rec2*
HD0319	TTTTGCGACATTGATCGCAAAA	HI0365
HD0182	TTTTGCGATCAAGATCGTGAAA	HI0938
	A. pleuropneumoniae	
Aple1014	TTTTGCGATCTTCATCGAAAAA	*comA*
Aple2116	TTTCTCGATCCTGATCGCAAAA	*comE1*
Aple1940	TTTTCCGATCCGTATCGCAAAA	*comF*
Aple1780	TTTTGCGATCCTGATCGAGAAA	*comM*
Aple1929	TTTTGTGATCTCAATCGAAAAA	*dprA*
Aple0139	TTTTGCGATACGGATCGCAGAA	*pilA*
Aple0635	TTTTGCGATCCGTGTCGAAAAA	*radC*
Aple0700	TTTTGCGATCAGGATCGAAGAA	*rec2*
Aple1575	TTTTGCGATCTTGATCGCAAAC	HI0365
Aple0828	TTTTGCGATCAAGATCGAATAA	HI0938
	M. haemolytica	
Not ann.	TTTTGCGATCCGCATCGAAAAA	*comA*
	no promoter sequence	*comE1*
	TTTTTCGAGCGATGTCACAAAA	*comF*
	no promoter sequence	*comM*
	TTTTGTGATCTCTCTCGAAAAG	*dprA*
	TTTTTCGATCTGCGTCGAAAAA	*pilA*
	TTTTGCGATCTTGCTCGAAAAA	*radC*
	TTTTGCGAACTGTGTCGAAAAT	*rec2*
	no promoter sequence	HI0366
	TTTTGCGATCTGCATCGAAAAA	HI0938

Table 2. Motif sites in *Pasteurellaceae* CRP-N-ortholog promoters.

Gene	Sequence	*H.i.* ortholog
	H. influenzae	
HI1434.1	TTTTGTGATCTACTTATCATTT	*cspD*
HI1615	TATTTTGCTTTGGCTAACATAA	*purE*
	AATTGTGCTTAGGATAAAATTT	
HI0745	TTATGTGATCGAGATCATAAAT	*ansB*
HI0287	TGATGTGAAAAATTCAATATTC	*mtr*
HI1350	ATAAGTGATCAAGATCACAGTT	*cdd*
HI0534	AAATGTGATCTTCATCAAGTTT	*aspA*
HI0131	AACTGTGAACTTCATCACGGTA	*afuA*
HI0835	TTTTTTGAGGTAGATCACAAAA	*frdA*
HI0610	AAGTGCGGTCGGTTTCACACCA	*fucR*
HI1356	ATTATTGACGAAGATCACACTT	*malQ*
HI1210	AAATGTGAACTAGATCATAGAA	*mdh*
HI0822	ATTTGTGACATGGATCACAAAT	*mglB*
HI0053	AACTGTGGCGTGGATCACAGTT	
HI0035	AAATGTGACGAACGTATCATTT	
HI1112	AACTGTGATCCACGCCACAGTT	*xylA*
HI1111	AACTGTGGCGTGGATCACAGTT	*xylF*
HI0815	AATTGTGATCTAGTACACAGTT	*uspA*
HI1662	GAGTTTGAACTAGATCACAAAT	*sucA*
HI0809	AAATGAGATCTACTTAACATTT	*pckA*
	ATTTTTGCTCTATATCACAATA	
HI1218	TTCTGTGATCCATCTCACAATC	*lctP*
HI0398	TTTTGTGACTCACTTCAAACTC	
HI0145	AAATGAGAAGTTGATCACATTT	
HI0146	AAATGTGATCAACTTCTCATTT	
HI1675	AATTATGATTTAAATCAATAAA	*moaC*
HI0608	TTTGTTGCTCTCGATCACATTT	
HI0590	TGGTGTGGTACAACTCACCATT	*potE*
HI0501	TTTTGTGATCAATATCCCAAAT	*rbsD*
HI0592	GTTTTTGACTAAGATCACATTT	
HI1227	TTAAATGAACAAGGTTACATTA	*uraA*
HI0740	AAATGTTAAGTAGATCAAAAAA	*yhxB*
HI0804	TTTTGTTAAACACTTCACATTT	
HI1124	TTATTAGACACAACTCACAAAA	*oppA*
HI0686	TTTTGTGATATTGATCACAATA	*glpT*
	ATTTGTGAAACACTTCACATTT	
HI1010	TTCTGTGATCTAGATCTCAGAT	
HI1645	TTTTGTGATAAAGATCTCATTC	*fbp*
HI1030	TAATATAAAACGAATCACATTT	
HI1031	AAATAGGATCTAGATCACAAAA	
HI1315	TTCTGTGATCCATCTCACAATC	
HI1126	ATTTGTGACTTGTATCACATTT	
HI0289	AAATTTTAACTTGATCACAATT	*sdaC*
	TTTTTTGCTTTGATTTACAATA	
HI1245	AATTGTGACGAACTGCAAACTT	
	M. succiniciproducens	
MS0956	TTATTTGAACAAGATCACAATT	HI0053
MS0698	TTTTGTTAACTTGATCACAATT	HI0053
MS1915	TTCTTTGAAGTAAATCACAAAT	HI0608
MS1583	ATTTGTGAACCATCTCACGGTA	*afuA*
MS2050	TCTTGTGAACTAGATCAAAAAA	*ansB*
MS1984	AAATTTGATTTAGATCACATTA	*aspA*

MS1095	TTTTGTGATCTCCGTTAAATTT	*cspD*
MS1615	AAATGTGCGTGAGATCACATTG	*fbp*
	AAATGATAGGTCTAACACAATA	
MS1652	TTTTTTGAGGTAGATCACAAAA	*sdhA*
MS1991	TATTGTGACTAAAATCACAAAT	*glpT*
MS0753	TTTTGTTAACTAAGTCACAATT	*lctP*
MS1124	TAATTTGAGTTAGATCACATAA	*malQ*
MS0643	TATTGTGAAAGCGATCACAGTA	*mglB*
MS1022	TTTTTTTATAAAAAACACATTA	*moaC*
MS2373	TTTTGTGATCTACGGCACAATT	*xylA*
MS0771	GCCTGAGAGATAAATCACAAAA	*yhxB*
MS1981	TTTTGTGATCTTTGTCTCAGTT	HI1010
MS0393	ATTTGTGGGTCAAAACTCATTA	HI1126
MS0349	ATTTTTGCCGATCATAACATAA	*uspA*
	AAACGTGATCTAGTGCAAATTT	
	P. multocida	
PM1071	AAATGTGATTACGGTTAAATTT	HI0035
PM1711	ATATGCGACAAAGATCTCAAAT	HI0145
PM1709	TTTTGTGACGAACCTATCATTT	HI0146
PM0805	AACTGTGATGGATATCACAAAT	HI0592
PM1167	TTTTATGCGCTTGTTCACAAAT	HI0608
PM0599	TAAGGTAATGAGGTTAACGTTT	HI0804
PM1366	TTTTGAGATCTCGATCGCAGAT	HI1010
PM1256	AAATGGGATCTTGATCACAAAA	HI1031
PM0002	TACTGTGTTTTAGGTCACGTTT	HI1245
PM0597	ATTTATGATCATGCTCATATTG	HI1315
	ATTTGTGATCTAACTCACCATG	
PM0953	TTTTGTGATAACTCTCACGGTA	*afuA*
PM0550	AAATTTGAGTTAGATCTCACTA	*mdh*
PM0156	TTATTTGATCCAGTTCACAGAT	*rbsD*
PM1103	AAGTGTTAACAGGATCAAATTA	*aspA*
	AAATGTGACGGCGATCAAATAT	
PM0481	TTTTGTGATCTCGGTTTGATTT	*cspD*
PM0930	AAATGTGTCGAAGATCACATTG	*fbp*
PM0201	TTTTTTGAGGTAGATCACGAAA	*frdA*
PM1443	AATTGTGACAGACATCACAAAT	*glpT*
	TTTTGTGAAATCACTCACAAAT	
PM1852	ATGTGTGAGTTTTGTCACAGAA	*lctP*
PM0540	TATCTTGACGAAGATCACTAAT	*malQ*
PM1038	AAGTGTGATCAAGGTAACAGTT	*mglB*
	AAATGTGAGTGAGATCACAGTC	
PM1192	AATTGCGTTGTTTAACAAAAAT	*mtr*
PM1910	AAAAATGATTTTCTCCACTTTT	*oppA*
	TTATCAAAAATAGCTCACAAAT	
PM1542	TTCTTTGACATAAATCATATAA	*pckA*
	AATTTTGATCAAGCTAACAGTT	
PM0619	AAATGTAGTTAGGATATGATTT	*purE*
PM0277	AAGTGCGACAGAGATCAAAAAA	*sucA*
	GTTAATGCTCTGTTACACAATT	
PM1286	AAACGTGATCTAAGGCATATTT	*uspA*
PM1074	CAAAGTGACTCAGTTCAAATAA	*yhxB*
	H. ducreyi	
HD0372	TTTTGTGAATAAGATCAAAGAA	*ansB*
HD0030	TTTATTGAGGTAGATCACAAAA	*frdA*
HD0264	GAATTTGCTTTATTTCACATTA	*mdh*
HD1428	AAGTTTGATTTATAGCAAATTT	*uspA*

	AAATGCGATCTAGTTCAAGTTT	
HD1331	TTTTTTGAAATTGATTATAATT	*pckA*
	AAATTTGAAGTACTTAATATTT	
	AAATATGATGAATATCATTTAA	
HD1852	AAATAGGATCTTAGTCACAATT	*nanE*
HD0868	TAATTTGAACTCCTTCACATTT	HI0608
HD1150	ACTTTTGAAAACGCTCACATTT	*glpT*
	AAATGTGGGGCATTTCACAATT	
HD0702	CATTGTGATCAATGTCACAAAA	*fbp*
HD0357	AATTTTGAAGTCATTCACATTT	HI1126
HD1143	AAACTCTAGCTAGATCACAAAA	*sdaC*

Table 3. Motif sites in *Enterobacteriaceae* CRP-S-ortholog promoters.

Gene	**Sequence**	***H.i.* ortholog**
	E. coli	
b3395	ATCTGCATCGGAATTTGCAGGC	*comA*
	TAAATCGAGCCTGCTCCCAGCA	
b0442	ATCCTGAAGCCGCCTCGCAAAA	*comE1*
	GCTTTCGCGGCCTTTTCCATTT	
b3413	AAATGCGAGCTAAGTTCCTCGT	*comF*
b3765	TTTTGCGAGCATCATTCCACCG	*comM*
b3286	CTTTGCGAAGCCGCTCGTCCGG	*dprA*
b0108	TTCTTCGTAACGCCTCGCAAAT	*pilA*
b3638	CTTTGCGAGGCGCTTTCCAGGA	*radC*
b0913	AACTGGAAGCTGCCTCGCAGAG	*rec2*
	ATATGCCTCGGGGAACGCAAAA	
b2826	TTCTTCGAGACGCCTTCCCGAA	HI0938
	S. typhimurium	
STM3492	ACCTGCATCGGAATTTGCAAAC	*comA*
	TAAATCGAGCCTGCTCCCAGCA	
STM0453	ATCGTCGAGGCGTTCGCAAAAA	*comE1*
	GCTTTCGCGGCCTTTTCCATTT	
STM3510	AAATGCGAGCCGAGTTCCTCGC	*comF*
STM3899	TTCTGCGAGCGTTCTTCCAGTT	*comM*
STM3405	CTTTGCGAAGGCGCTCGTCCGG	*dprA*
STM0144	ATATTCGTAGCGCCTCGCAATA	*pilA*
STM3729	CTTTGCGAGGCGCTACGCAAGA	*radC*
STM0983	AACTGGAAAACGTTTCGCATTT	*rec2*
STM4256	ACCTGGAACCTGCATCGCAGCT	*ssb*
STM3000	ATCTTCGGCGCGCATTCCTGAA	HI0938
	Y. pestis	
y3925	ACCTGCATAGGTGTTTGCAGCC	*comA*
	AATATCGAGGCTGCTCCCAGTA	
y1032	AACCGCAATAAGCTTCGCATTC	*comE1*
	GCTTTGGCGACCTTTCGCATAT	
y0334	TTTTGCATACCTCATCGCAGTT	*comM*
	TTTCTCGTGAGCTTTCGCAAAC	
y4024	TTTTGCGCAGCCGTTCGTCTGG	*dprA*
y0761	TCCGTCAATACGCCCCGCAATT	*pilA*
	TTTTGCGAGTGCCGCCGAAGTT	
y0092	ATTTGCGAGACGTCACGCATGC	*radC*
y2778	GGTTTCGATACATCCCGCATTT	*rec2*
y0582	AACTGCAATATATTTCGCAGTT	*ssb*

y3170	AAATGCGAGTCGTATCGCAGAC TTTTGCGTACCGCTTCCAACAC	HI0938

Table 4. Motif sites in *Enterobacteriaceae* CRP-N-ortholog promoters.

Gene	Sequence	*H.i.* ortholog
	E. coli	
b3685	CATATTGATTTAATTCGTAATG	HI0035
b2736	TTATGTGAATCAGATCACCATA	HI1010
b3577	AATTGTGGTTAAAGTCGCATTA	HI1030
b3575	AAGTGTGCCGTAGTTCACGATC	HI1031
b2463	ATGAGTGCGTTAATTCACACTT	HI1245
b3679	AATTCCGCTGGAGATCACATTT	HI1315
b2143	ATTTGCGATGCGTCGCGCATTT TAATGAGATTCAGATCACATAT	*cdd*
b0880	ATCAGCGACATCTGTCACATTC	*cspD*
b2240	TTGTTTGATTTCGCGCATATTC AAACGTGATTTCATGCGTCATT	*glpT*
	ATGTGTGCGGCAATTCACATTT	
b3417	TTCTGCGCTGTATTGCATTGAT TTAAGTGGTTGAGATCACATTT	*malQ*
b2150	ATCTGTGAGTGATTTCACAGTA	*mglB*
b3403	GAATGCGATTCCACTCACAATA ATCTATGAGCCTTGTCGCGGTT	*pckA*
b2796	ATTTGAGATCAAGATCACTGAT	*sdaC*
b3565	ATTTATGACCGAGATCTTACTT TTTTGCGAGCGAGCGCACACTT	*xylA*
b3566	AAGTGTGCGCTCGCTCGCAAAA AAGTAAGATCTCGGTCATAAAT	*xylF*
b2801	TAAAGTGATGGTAGTCACATAA AAGTGTGACCGCCGTCATATTA	*fucP*
b3603	ATCTGACCTCTGGTTCACAATT	*lctP*
b3748	CGTTTCGAGGTTGATCACATTT	*rbsD*
	S. typhimurium	
STM2183	ATTTGCGATACGTCGCGCATTT CAATGAGATTTAGATCACATAT	*cdd*
STM2970	ATTTGAGATCGGGATCACTGAT	*sdaC*
STM3881	CGTTTCGACGGCGATCACAATT	*rbsD*
STM0943	ATCCGCGACATCTGTCACATTC	*cspD*
STM3668	AAGTGTGTTGCAGTTCACGATA	HI1031
STM2283	ATGTTTGATTTCGCGCATAATC AAACGTGATTTCGTGCGCCTTT	*glpT*
	ATATGTGCTGTAATTCACATTA	
STM2974	TTAATTGATGTGAATCACAAAA	*fucA*
STM2472	ATGAGTGTGTTGATTCACACTT	HI1215
STM3661	GGATTCGATCGCGATCGCTTTT TTTTGAGAGCCAGAGCACATTT	*xylA*
STM3514	GTAAGTGGCGGCGATCACACTT	*malP*
STM3500	GAATGCGATTACAGTCACATTA CTGCGTGACAGGAGTCACAGTG ATCTATGAGCCTTGTCGCGGTT	*pckA*
	Y. pestis	
y2657	TAATGAGATATAAATCACAATT	*cdd*
y2862	AATTGAGATCACGATCACGGTA	*sdaC*

y2662	ATTTGTGGTGTTGCTCACTCGT	*mglB*
	ATCTGTGAGAAAATTCACAGTT	
y0007	TGTTTCGGTGGCGATCACAATT	*rbsD*
y4100	TTTTGTGGCGTATCCCACATTC	HI0035
y3859	TTGAGTGTTTGCTTACACATTA	*uspA*
y2787	TTGCGTCATTGTCTTCACTTTT	*ansB*
y4057	TTATGAGATCTACACCACAATT	*xylA*
y4056	AATTGTGGTGTAGATCTCATAA	*xylF*
y3918	ATTCGTGTTCCATCTCTCATAA	*pckA*
	ATATTTGATAGCTATCGCTGTT	
y0668	TAATGTGCGCTATCTCATTAAT	*mdh*
	TATTGTGTTTAAAATCACAATA	

Table 5. Motif sites in *Vibrionaceae* CRP-S-ortholog promoters.

Gene	**Sequence**	***H.i.* ortholog**
	V. cholerae	
VC2634	AAGATTGTAGTGACTCCAAGAA	*comA*
	CTTTATGAACTTCACCGGAGAA	
VC0047-8	AATATCGACTTGGGTCGCCGCT	*none-dprA*
	TTGTTCGACCGGTTTCGCAACG	
	ACAGACATATACACTCGAAATG	
VC2423	AGTTTTTAACTGACTCGAAGTT	*pilA*
VC1612	ATTTGCCAACTGACTCGCAGAC	HI0366
VCA0140	GAGTTTGAAGTGCCTCGAAGAG	*None*
	V. parahaemolyticus	
VP2750	AAAATTGTGGTGACTCCAAGAA	*comA*
	CTTTATGAACTTCACCGGAGAA	
VP3041-0	GATATCAACCTGCGTTGCAGCA	*none-dprA*
	AATATTGAACTGTGCCGAAACA	
	ACAGACATATTCACTCGAAATA	
VP2523	GAGTTTTACCTCACTCGAGACC	*pilA*
VP1752	GTTTGCAAACCTGATCGCATAG	HI0366
VPA0092	GAATTTGAAGTGACTCGAAAGA	*None*
	V. vulnificus	
VV2994	AAAATTGTTGTGACTCCAAGAA	*comA*
	CTTTATGAACTTCACCGGAGAA	
VV3224-3	CTTTTCAACCGGTTTGGCTACT	*none-dprA*
	AATGTTGAACTGTGCCGAAACA	
	ACAGACATATGCACTCGAAATG	
VV2778	GAATTTTAAATCACTCGAGTGA	*pilA*
VV1491	TTATGCAAAGTGACTCGCATTG	HI0366
VVA0086	CAATTTGAAGTGACTCGAAAAT	*None*

Table 6. Motif sites in *Vibrionaceae* CRP-N-ortholog promoters.

Gene	**Sequence**	***H.i.* ortholog**
	V. cholerae	
VC1231	ATGTGTGACGTCACTCTAATAA	*cdd*
	TAACGTGACACTGATCACCTTA	
VC1779	AATTTTGTTCGCCATCACACTT	HI0146

VC1325	TGGTGTAAACGTTATCACTCAT	*mglB*
	AATTGTTATTGAGTTCAAACTA	
	ATTTTTTAACTGGTTCACATTA	
VC2656	AATTGTGACACCAGTCACATAT	*frdA*
	GTCAGTGAGTTCCATCTCAGTA	
VCA0160	TTTTTTGACCTGAAAAACATAA	*mtr*
VC1781	AAGTGTGATGGCGAACAAAATT	HI0145
VCA0013	TGGTGTGATCCGAATCACTGCT	*malP*
	TGATGGGAGCTAGATCACTCAC	
VC0052	TTTGGTTATCCGGATCACACCC	*purE*
VC2738	ATATTTGAGCTGCCTCCCTGTT	*pckA*
	V. parahaemolyticus	
VP1298	ATGTGTGAGCTCACTCTAATTA	*cdd*
	TAACGTGATCTAAAGCACGGAA	
VPA1702	ATTTGTGTAGGGTCTCAAAATA	HI0146
	TCACGTGAGCAGCTTCACAAAT	
	TAGTGTGATTTTGGTCAATCAA	
VPA1067	TGATGTGATAACAATCACTAAA	*rbsD*
VP2840	TGCCGTGATAGCAGTCACATAA	*frdA*
VPA0374	TGTTGAGCTTGTGCTCAAAAAT	*ansB*
VPA1620	GGTTGTGATCAAAATCACTAAG	*malP*
	TGTTGAGATTTGGATCACTAAC	
VP0129	TTTTGTGATCTATCCCCCGTAA	*pckA*
	ATTTTTGAACTATCTCCCTGTT	
VP0325	TTCTGTTAGTTGCATCACTGTA	*mdh*
	TTAATTGATTGTAATCAAGTTG	
	V. vulnificus	
VV0434	TGCTGTTTTTTCAGTCACTTTT	*fbp*
	TCAAGAGATGCCGCTCACACTC	
VV1962	ATTTGTGACATCACTCTAATAA	*cdd*
	TATAGTGACAGAGATCACTGAA	
VVA1590	ATTTGTGTAGGGTCTCAAAATA	HI0052
	TCACGTGAGCGGCTTCACAAAT	
	TAGTGTGATTTTGGTCAATCAA	
VVA0544	CGATGTGATATATATCTCTAAA	*sdaC*
VVA0163	AATTTTTATCTAGTTCACATTA	*mglB*
VV3097	TGCCGTGACAGTTATCACATAA	*frdA*
VVA0568	ATATGGGACAAAAGTAACGTAA	*rbsD*
VVA0966	AAATGTAACATTTCTCACAGAA	*glpT*
VVA1204	AAATGTGATCGCGAACAGAAAT	HI0145
VV3010	AAGATTGACTTATATCAATTAG	HI1245
VVA0077	TATTGTGATCGAATTTACAAAA	*malP*
	GGCTGTGATCTCAATCACTGCA	
	AGGTGAGAGACGGATCACTAAC	
VV0207	ATTTTTGAACTATATCCCTGTT	*pckA*
VV0467	ATCTGTTAGTTGTATCACTGTA	*mdh*
	TTAATTGATTGTAATCAAGTTG	
	ATAAGAGATCGCTCTCAAGGAG	

Table 7. Motif sites in *Vibrionaceae* CRP-N-ortholog promoters that resemble PurR sites.

Gene	**Sequence**	***H.i.* ortholog**
	V. cholerae	
VC2544	GCGCAATCGATTCCAT	*fbp*

VC1231	TTGCAATCGTTATCAT	*cdd*
VC1325	GTGTAAACGTTATCAC	*mglB*
	GAGTAAACGTTTTCAC	
VCA0127	ATCGAAACGTTTCGAT	*rbsD*
VC2171	TCGCAATCGATTGCAG	*uraA*
VC0052	AAGCAAACGTTTGCTT	*purE*
VC2738	GCGCAAAGGTTTGCGC	*pckA*
VC0432	TCGCATACCTATGCAT	*mdh*
	V. parahaemolyticus	
VP0313	GCGCAAACGTTTAACA	*fbp*
VP1298	TTGCAATCGAATACAT	*cdd*
VPA1087	ATCGAAACGTTTCGAT	*rbsD*
	ATCGAAACGTTTCGAT	
VP2283	TTGCAAACGATTGCAG	*uraA*
VP2019	ATCGAAAGTTTTGGCT	*oppA*
VP3036	ACGCAAACGTTTGCTT	*purE*
VP0129	GCGCAAAGGATTGCGC	*pckA*
	V. vulnificus	
VV0434	GCGCAAACGATAACCT	*fbp*
VV1962	GTGCAATCGAATACAT	*cdd*
VVA0163	GGGTAAACGTTTTCAC	*mglB*
VVA0568	ATCGAAACGTTTCGAT	*rbsD*
	CGCGAATCGATTGAGT	
VV2324	AGGCTAAAGATTGGCT	*cspD*
VVA0966	ACGAAACCGTTTGCTC	*glpT*
VV2513	TAGCAATCGTTTGCAA	*uraA*
VV3218	ACGCAAACGTTTGCTT	*purE*
VV0207	GCGCAAAGGTTTGCGT	*pckA*