J. theor. Biol. (2000) **207,** 000–000
doi:10.1006/jtbi.2000.2179, available online at http://www.idealibrary.com on IDEAL®

Regulation of Competence Development in *Haemophilus influenzae*

Proposed Competence Regulatory Elements are CRP-Binding Sites

LEAH P. MACFADYEN*

Department of Zoology, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z4

(*Received on* 21 *October* 1999, *Accepted in revised form on* 24 *August* 2000)

Development of competence for DNA uptake by the bacterium *Haemophilus influenzae* is tightly regulated, and expression of the cell's complement of competence genes is absolutely dependent on the cAMP–CRP complex. A second regulator of competence may maximize competence under starvation conditions. Several investigators have recently identified a consensus sequence (competence regulatory element, CRE) in the promoter regions of some competence genes and have proposed that this may be a binding site for Sxy (TfoX), a putative positive regulator of competence. However, a scoring method that reliably ranks candidate binding sites according to affinity for the cognate binding protein predicts that the cAMP–CRP complex will bind CRE sequences with high affinity. Moreover, the predicted Sxy protein lacks recognizable DNA-binding motifs and has not been shown to bind DNA. No other consensus sequences (putative binding sites) were identified in the promoter regions of competence genes. These observations suggest that the proposed competence-specific regulatory elements are in fact CRP-binding sites, and highlight the central role of cAMP—an established bacterial mediator of the response to nutritional stress—in competence regulation. Minor sequence elements uniquely conserved in the set of CRE sequences are predicted to reduce CRP affinity, and a model is suggested in which a secondary regulator of competence genes may interact with CRP under certain conditions to stabilize the initiation complex.

Introduction

REGULATION OF COMPETENCE DEVELOPMENT BY *HAEMOPHILUS INFLUENZAE*

DNA uptake by H. influenzae

Haemophilus influenzae is a small Gram-negative bacterium capable of developing natural competence for DNA uptake. When competent, *H. influenzae* cells can bind several hundred kilobases of free DNA and transport it into the cytoplasm. Some of the transported DNA may be incorporated into the chromosome by homologous recombination (Setlow *et al.*, 1981, 1988), the remainder is degraded and the nucleotides recycled (Pifer & Smith, 1985). We are investigating the regulation of natural competence in this organism, because systems regulating competence development probably evolved to maximize the benefits and minimize the costs of DNA uptake. An understanding of the nature of signals and mechanisms regulating expression of competence genes should illuminate the primary benefits of this process.

* Correspondence address: Suite 314-2255 West 4th Ave., Vancouver, British Columbia, Canada V6K 1N9.
E-mail: macfad@uwc.net

0022–5193/00/000000 + 00 $35.00/0

Competence Development is Dependent on cAMP–CRP and on the Sxy (TfoX) Regulator

Expression of a number of competence genes (Tomb *et al*., 1991; Larson & Goodgal, 1991; Clifton *et al*., 1994; Karudapuram *et al*., 1995; Gwinn *et al*., 1997, 1998; Karudapuram & Barcak, 1997) and the development of competence are tightly regulated. Competence is low during exponential growth in rich medium, and develops spontaneously at the onset of stationary phase, or when cells are transferred to a nutrient-limited medium (Herriott *et al*., 1970). Competence development in *H. influenzae* is absolutely dependent on the cyclic nucleotide 3′,5′-cyclic adenosine monophosphate (cAMP) (and on its receptor protein, CRP (Chandler, 1992). Cyclic AMP is a central mediator of the response of the enteric bacteria to nutritional stress, and intracellular levels of cAMP rise on entry into stationary phase of growth (Peterkofsky & Gazdar, 1971), or after transfer of cells to a nutrient-limited environment (Death & Ferenci, 1993; Notley & Ferenci, 1995). A complex of cAMP with CRP has been shown to bind conserved sequences in promoter regions of cAMP-regulated genes and activate their transcription (Botsford & Harman, 1992). An *H. influenzae* mutant strain lacking adenylate cyclase—the enzyme responsible for synthesis of cAMP—is completely competence-deficient, but competence is restored to this *cya* strain by addition of exogenous cAMP (Dorocicz *et al*., 1993). Cells lacking the cAMP receptor protein (CRP) homologue are also completely competence deficient (Chandler, 1992). It is believed, therefore, that a cAMP–CRP complex also acts as a transcriptional regulator in *H. influenzae*. Because the DNA-binding domains of *E. coli* and *H. influenzae* CRP orthologues are very similar, it was predicted that the promoter region of one or more competence genes would contain sequences similar to the *Escherichia coli* CRP consensus binding site.

Competence development is also believed to be positively regulated by the product of the *sxy* (*tfoX*) gene (Redfield, 1991; Williams *et al*., 1994; Zulty & Barcak, 1995). Disruption of *sxy* completely prevents competence development (Williams *et al*., 1994; Zulty & Barcak, 1995). Expression of the *dprABC* operon (proposed to encode enzymes for processing of transported DNA) (Karudapuram & Barcak, 1997), the late competence gene *com101A* (*comF*) (Zulty & Barcak, 1995) and possibly the entire *com* operon is absolutely dependent on expression of the Sxy protein.

Conserved Promoter Sequence Elements are Proposed to be Binding Sites for Sxy

Recently, several investigators have identified a conserved 26 bp palindromic "competence regulatory element" (CRE; also known as a "dyad symmetry element", DSE) in the promoter regions of a number of genes required for and induced during competence development (Table 1). The CRE was shown to be required for induction of the *com* operon, and it has been proposed that this sequence is a binding site for the Sxy protein (Tomb *et al*., 1993; Karudapuram & Barcak, 1997). Moreover, the *sxy* promoter possesses a candidate CRP-binding site, and Zulty & Barcak (1995) demonstrated a three-fold induction of Sxy expression after addition of 1 mM cAMP. A simple cascade model of competence induction has therefore been proposed in which increased levels of cAMP trigger expression of Sxy, and Sxy then activates transcription of competence genes, including the *com* operon (Tomb *et al*., 1993; Karudapuram & Barcak, 1997).

Problem: *CRE Sequences Resemble CRP-binding Sites*

The published CRE consensus closely resembles the CRP consensus binding site:

CRP: WWW<u>TGTGA</u>TNTANA <u>TCACA</u>WWW,

CRE: ATTTT<u>TGCGA</u>TCYGCA<u>TCGCA</u> AAATT,

where W = A or T, and Y = C or T. Like CRP-binding sites (Botsford & Harman, 1992), CRE sequences are positioned upstream from suboptimal promoters, and at a distance roughly equal to an integral number of turns of the DNA helix from the proposed RNA-polymerase-binding site

TABLE 1

Sequences identified as competence regulatory elements (CRE) in the promoter regions of genes known or expected to be required for DNA uptake and/or recombination

HI #*	Gene	Function	Putative CRE sequence	Reference†	Score (I_{seq})‡
0061	*rec-2*	Competence (DNA translocation?)	TTTTGCGATCCATATCGTAAAA	(Karudapuram & Barcak, 1997)	12.99
0299	*pilA*	Fimbriae; competence	TTTTGCGATCAGGATCGCAGAA	(Dougherty & Smith, 1999)	16.30
0439	*comA*	Competence	TTTTGCGATCCGCATCGTAAAA	(Karudapuram & Barcak, 1997)	14.23
0985	*dprA*	Competence (DNA translocation?)	TTTTGCGATCTGCATCGCAAAA	(Karudapuram & Barcak, 1997)	16.84
1008	*comE1*	Homologue of *B. subtilis* competence gene	TTTTGCGATCGAGATCGCAAAA	(Karudapuram & Barcak, 1997)	16.05
1117	*comM*	Competence	TTTTGCGATCTAGATCGCAAAA	(Gwinn *et al.*, 1998)	18.05
0250	*ssb*	Recombination/DNA repair (single-stranded DNA-binding protein)	TTTTGCGATCATTATCGCATAT	(Macfadyen, 1999)	16.83
0364	*pbp7*	Cell envelope biosynthesis	TTTTGCGATCTAGATCGCAAAT	(Karudapuram & Barcak, 1997)	18.05
1181	*gmhA*	Cell envelope biosynthesis	TTTTGCGATTTAGATCGAAAAA	(Karudapuram & Barcak, 1997)	16.09

*Gene numbers for *H. influenzae* genes (HI #) are as assigned by Fleischmann *et al.* (1995) and The Institute for Genome Research (1999).
†References are given for the original identification of each binding site.
‡Total information content for the putative binding site, calculated using the matrix shown in Table 2.

(Karudapuram & Barcak, 1997). I have now employed an established method of scoring candidate binding sites of DNA-binding proteins to assess the likelihood that the cAMP–CRP complex will recognize and bind these regulatory sequences. Results suggest that CRE regulatory elements are, in fact, CRP-binding sites.

Theory

PREDICTING AFFINITY OF A BINDING PROTEIN FOR A CANDIDATE BINDING SITE

The affinity of CRP for a binding site has been shown to depend on the degree to which the sequence resembles the CRP-binding site consensus sequence [see Botsford & Harman (1992) and references therein]. However, some positions in CRP-binding sites are more conserved than others, and so any attempt to score CRP affinity for a candidate CRP site by simply counting the number of matches to the consensus sequence will be inaccurate. A more accurate method of scoring candidate binding site sequences such as CRP-binding sites was developed by Stormo & Hartzell (1989), and incorporates a weighting for more conserved positions in the binding site sequence. This method calculates a goodness-of-fit score for each base at each position of the candidate binding site, using an aligned set of known binding site sequences, and gives a matrix representation of the binding site, as follows.

In classical statistical analysis, the discrepancy between the observed frequency and the expected frequency of any result (or "likelihood") is expressed as the ratio of their frequencies:

$$\left(\frac{f_i}{\hat{f}}\right),$$

where f_i represents observed frequency and $\hat{f}$ represents expected frequency. The ratio of these two frequencies can be used as a statistic to measure the degree of agreement between sampled and expected frequencies. Goodness of fit, G, is calculated as

$$G = 2\ln\left(\frac{f_i}{\hat{f}}\right),$$

and the greater the departure from expectation, the greater is the value of G (Sokal & Rohlf, 1969). Stormo & Hartzell recognized that the degree to which choice of base is constrained at a given position, b, in the DNA sequence of a protein-binding site can be expressed as a measure of goodness of fit or "information content", I_b:

$$I_b = \log_2\left(\frac{f_b}{p_b}\right),$$

where f_b is the observed frequency of occurrence of a base at a given position in a set of aligned sequences known to be binding sites, and p_b represents the predicted frequency of occurrence of this base, in the absence of constraints, calculated from the base composition of the organism. The more frequently a given base occurs at a given position in the protein-binding site, the greater is the value of I_b for that base. Moreover, using a matrix constructed from aligned sequences of known binding sites, an information score can be calculated for any candidate binding site sequence as

$$I_{seq} = \sum^{n} I_b,$$

where I_b represents the information content score for each base in the sequence, and n is the number of bases in the sequence. An I_{seq} score is therefore a weighted measure of how closely a specific sequence fits the constraints thought to be imposed by its function, and these scores can be used to rank the affinities of a binding protein for different candidate binding sites. Such scores have been shown to do well as predictors of quantitative activity, when compared to experimental data (Mulligan *et al.*, 1984; Berg & von Hippel, 1987, 1988; Stormo, 1988).

Method

PREDICTING AFFINITY OF CRP FOR PROPOSED COMPETENCE REGULATORY ELEMENTS

In 1989, Stormo & Hartzell constructed a matrix for the *Escherichia coli* CRP-binding site, using the available set of 23 characterized CRP-binding site sequences. The accuracy with which

TABLE 2
Matrix representation of the Escherichia coli CRP-binding site

		Position in binding site (bp #)																					
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
A	G	4	6	9	2	40.	0.5	41	3	9	10	5	17	21	7	7	0.5	10	0.5	6	7	0.5	4
	A	23	23	18	0.5	2	1	4	43	6	9	9	15	10	16	6	7	35	10	33	12	18	14
	T	21	18	18	39	6	40	4	0.5	24	14	20	10	10	15	33	3	3	7	6	21	30	24
	C	1	2	4	8	1	8	0.5	3	10	17	15	7	8	11	3	39	1	32	4	9	1	7
B	G	0.08	0.12	0.18	0.04	0.82	0.01	0.84	0.06	0.18	0.2	0.1	0.35	0.43	0.14	0.14	0.01	0.2	0.01	0.12	0.14	0.01	0.08
	A	0.47	0.47	0.37	0.01	0.04	0.02	0.08	0.88	0.12	0.18	0.18	0.31	0.2	0.33	0.12	0.14	0.71	0.2	0.67	0.24	0.37	0.29
	T	0.43	0.37	0.37	0.8	0.12	0.82	0.08	0.01	0.49	0.29	0.41	0.2	0.2	0.31	0.67	0.06	0.06	0.14	0.12	0.43	0.61	0.49
	C	0.02	0.04	0.08	0.16	0.02	0.16	0.01	0.06	0.2	0.35	0.31	0.14	0.16	0.22	0.06	0.8	0.02	0.65	0.08	0.18	0.02	0.14
C	G	−1.61	−1.03	−0.44	−2.63	1.71	−4.63	1.73	−2.04	−0.44	−0.29	−1.29	0.47	0.78	−0.81	−0.81	−4.63	−0.29	−4.63	−1.03	−0.81	−4.63	−1.61
	A	0.91	0.91	0.56	−4.63	−2.61	−3.63	−1.63	1.80	−1.03	−0.44	−0.44	0.29	−0.29	0.39	−1.03	−0.82	1.51	−0.31	1.43	−0.03	0.54	0.19
	T	0.78	0.56	0.56	1.66	−1.03	1.69	−1.63	−4.63	0.97	0.19	0.71	−0.29	−0.29	0.29	1.43	−2.04	−2.03	−0.82	−1.03	0.78	1.28	0.97
	C	−3.61	−2.61	−1.61	−0.63	−3.61	−0.63	−4.63	−2.04	−0.29	0.47	0.29	−0.81	−0.61	−0.16	−2.03	1.66	−3.61	1.37	−1.61	−0.44	−3.63	−0.81

Note: A—No. (n) of occurrences of each base at each position from a set of 49 characterized *E. coli* CRP-binding sites (Robison & Church, 1994). Where occurrence = 0, an estimated occurrence of 0.5 was used. B—The frequency with which each base occurs at each position in the CRP-binding sites ($f_b = n_b/49$). C—Specifity matrix for CRP, based on the binding sites, which is calculated as $\log_2(f_b/p_b)$, where f_b is the observed frequency of each base (from the matrix above) and p_b is the *a priori* probability of obtaining base b. Here, $p_b = 0.25$ for all b, approximating the *E. coli* genomic composition.

TABLE 3
Putative CRP-binding sites identified in the promoter regions of selected Haemophilus influenzae genes

HI#	Gene*	Function	Putative CRP site sequence†	Score (I_{seq})
0604	*cya*	Adenylate cyclase	AATTGTGATTTATGTCACATTT	22.44
0398	hyp.	Promoter proximal gene of *icc* operon	TTTTGTGACTCACTTCAAACTC	16.38
0957	*crp*	cAMP receptor protein	AAGCGTGATTTTACGCGAAGGA	5.23
0185	*adhC*	Alcohol dehydrogenase	TTTTGTGATATGGCTCACAAAA	20.90
1739.1	*lctD*	Lactate utilization (L-lactate dehydrogenase)	AATTGTGATCTAGTTCTCAAAA	19.03
0851	*mobB*	Dinucleotide biosynthesis protein B	TACTGCGATTTAGATCGCAAAC	14.95
0937	*suhB*	Role in protein synthesis	TTTMGCGATCTGTATCGCAAAG	13.07
1112	*xylA*	Xylose isomerase	AACTGTGGCGTGGATCACAGTT	15.54
0822	*mglB*	D-Galactose binding protein	ATTTGTGACATGGATCACAAAT	21.09
0501	*rbsD*	Ribose transport protein	TTTTGTGATCAATATCCCAAAT	15.60
0615	*fucR*	L-Fucose operon activator	TTTTGTGAGTTTCTTTTCAAGA	6.00

cya was cloned and sequenced by Dorocicz *et al.* (1993) and *crp* was cloned and sequenced by Chandler (1992). All other genes were sequenced and identified by Fleischmann *et al.* (1995). hyp., hypothetical open reading frame.
†Putative CRP sites for *cya* (Dorocicz *et al.*, 1993) and *crp* (Chandler, 1992) were identified by sequence gazing, inspired by the existence of CRP binding sites in the promoter regions of their *E. coli* homologues (as reviewed by Botsford & Harman, 1992). Putative CRP sites for *xylA*, *mglB*, *rbsD* and *fucR* were identified by sequence gazing by Macfadyen *et al.* (1996), after determination of cAMP–CRP-dependent sugar utilization phenotypes. Putative CRP site for HI#0398 (Macfadyen *et al.*, 1998) was identified by sequence gazing. All other listed putative CRP-binding sites were identified by Dr R. J. Redfield, by BLAST searching of the *H. influenzae* genome for sequences similar to the *E. coli* CRP-binding site consensus sequence (Botsford & Harman, 1992).

a matrix represents a conserved binding site sequence increases with the number of aligned sites used in its construction. I therefore used the method described above to construct a new, more accurate matrix (Table 2) for the CRP binding site using the 49 characterized *E. coli* CRP-binding sites now listed in the DPInteract database (Robison & Church, 1994).

I used this matrix to score five sets of sequences: a set ($n = 100$) of randomly selected 22 bp sequences from the *H. influenzae* genome (Fleischmann *et al.*, 1995; TIGR, 1999), a set ($n = 100$) of randomly selected 22 bp sequences from the *E. coli* genome (Blattner *et al.*, 1997; GenBank, 1999), the available set ($n = 49$) of characterized *E. coli* CRP sites (Robison & Church, 1994), a set ($n = 11$) of candidate *H. influenzae* CRP sites from promoter regions of catabolic and biosynthetic genes (Table 3) and the set ($n = 9$) of putative *H. influenzae* CRE regulatory sites found in the promoter regions of established or putative competence genes (Table 1). Distribution of I_{seq} scores for each set of sequences is shown in Fig. 1.

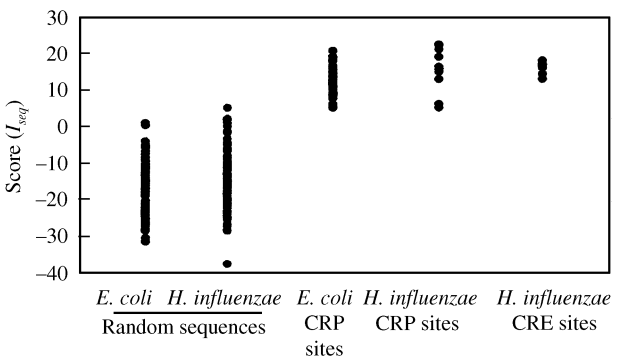

FIG. 1. Scatter plot showing distribution of I_{seq} scores for *Escherichia coli* and *Haemophilus influenzae* sequences. I_{seq} scores, representing relative affinity of CRP for each sequence, were calculated using the matrix shown in Table 2.

I then applied statistical methods [using JMP IN® Version 3.2.1 (SAS Institute Inc.) and Microsoft® Excel 98 (Microsoft Corporation)] to determine whether there is significant difference or similarity between any of these groups of scores. Shapiro–Wilk tests (Zar, 1996) determined that all five sets of scores are normally distributed, but score variances for each set were found to differ significantly (not shown). These five samples also differ in size, precluding the use of a simple ANOVA comparison of mean scores. Instead, I used the non-parametric

Kruskal–Wallis test (Zar, 1996) to test the null hypothesis that these sets of scores do not differ significantly. Because this test rejected the null hypothesis (not shown), I subsequently compared pairs of samples of interest using a non-parametric test developed by Dunn (Zar, 1996) (Table 4).

TABLE 4
Dunn's non-parametric pairwise comparisons of sets of I_{seq} scores

Sample comparison	Difference $(\bar{R}_A - \bar{R}_B)$	S.E.	Q	$Q_{0.05,4}$	Conclusion
E. coli random vs. *H. influenzae* random	17.87	11.00	1.6241	2.807	Same
H. influenzae random vs. *E. coli* CRP sites	120.78	13.57	8.903	2.807	Different
E. coli random vs. *E. coli* CRP sites	138.65	13.57	10.220	2.807	Different
E. coli random vs. *H. influenzae* CRE sites	152.25	24.71	6.16	2.807	Different
H. influenzae random vs. *H. influenzae* CRE sites	133.34	27.27	4.888	2.807	Different
H. influenzae random vs. *H. influenzae* CRP sites (putative)	134.38	24.71	5.438	2.807	Different
E. coli CRP sites vs. *H. influenzae* CRP sites (putative)	13.60	25.96	0.524	2.807	Same
H. influenzae CRE sites vs. *E. coli* CRP sites	20.06	28.21	0.711	2.807	Same
H. influenzae CRE sites vs. *H. influenzae* CRP sites (putative)	6.46	34.97	0.185	2.807	Same

Results

MATRIX-DERIVED SCORES REFLECT CRP AFFINITY FOR CRP-BINDING SITES IN BOTH *E. COLI* AND *H. INFLUENZAE*

In order to test the prediction that matrix-derived scores for candidate CRP-binding sites reflect the relative affinity of CRP for these different sequences, I plotted I_{seq} scores for a set of known *E. coli* CRP-binding sites (Table 5) against their experimentally determined CRP-binding affinities. Figure 2 shows that experimentally determined CRP affinity is roughly proportional to the calculated I_{seq} score for each site, confirming that the matrix-derived I_{seq} score reflects the utility of each site as a CRP-binding site.

It was also necessary to determine whether the *E. coli*-derived CRP matrix could provide meaningful I_{seq} scores for *H. influenzae* sequences, despite the difference in base pair composition between these two organisms (50% GC in *E. coli*, 38% GC in *H. influenzae*). This difference will necessarily affect the degrees of constraint on individual bases occurring at given positions in DNA-binding site sequences. In the absence of a set of experimentally confirmed *H. influenzae* CRP-binding site sequences, it is, however, impossible to construct a species-specific matrix. Both the mean matrix-derived I_{seq} scores (-16.51 vs. -14.22) and the score distributions (Fig. 1) of sets of 100 randomly selected 22 bp sequences from *E. coli* and *H. influenzae* differ slightly, presumably as a result of differences in genomic base composition. However, scores for randomly selected samples of *E. coli* and *H. influenzae* sequences are not significantly different ($p > 0.05$) (Table 4). This suggests that the *E. coli* matrix will generate sufficiently accurate CRP-affinity scores for *H. influenzae* sequences for the purpose of this study.

TABLE 5

Escherichia coli CRP-binding sites with known affinities for CRP

	Function	CRP-binding site sequence (de Combrugghe *et al.*, 1984)	Affinity ($\ln(K_s/K_{ns})$)*	Score (I_{seq})†
lacZ$_1$	Lactose utilization (□-galactosidase)	TAATGTGAGTTAGCTCACTCAT	9.0	18.30
lacZ$_2$	Lactose utilization (□-galactosidase)	AATTGTGAGCGGATAACAATTT	5.0	8.94
galE	Galactose utilization (UDP-galactose-4-epimerase)	AAGTGTGACATGGAATAAATTA	6.5	13.47
malT	Maltose utilization (regulator of *mal* regulon)	AATTGTGACGCCGTGCAAATAA	8.2	16.71
uxuA	Hexose utilization (mannonate hydrolase)	TGTTGTGATGTGGTTAACCCAA	7.2	14.08

*Affinity is represented as difference in binding for the specific sequence and a non-specific site (Berg & von Hippel, 1988).
†Total information content for the putative binding site, calculated using the matrix shown in Table 2.

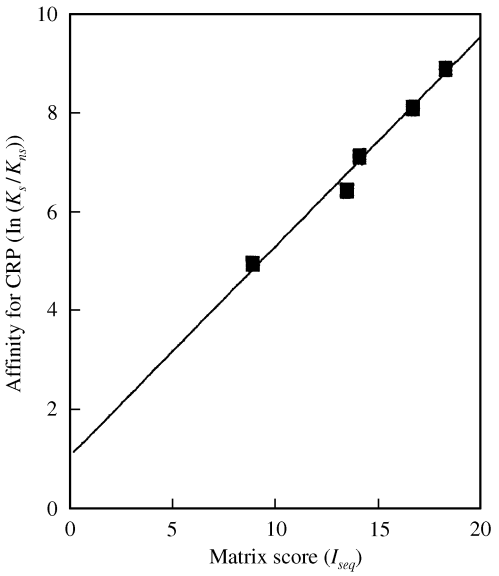

FIG. 2. Correlation between theoretical scores of CRP-binding site affinity (I_{seq}) and experimentally determined CRP affinity ($\ln(K_s/K_{ns})$) for known *E. coli* CRP-binding sites. Scores for the five *E. coli* CRP-binding sites listed in Table 5 were calculated using the matrix shown in Table 2. Experimentally determined values for CRP-binding affinity were taken from Berg & von Hippel (1988). Line was fitted by least-squares analysis using CricketGraph IIITM Version 1.01.

DERIVED CRP-AFFINITY SCORES FOR CRE SEQUENCES DO NOT DIFFER SIGNIFICANTLY FROM SCORES OF CHARACTERIZED CRP-BINDING SITES

Having established the efficacy of this method for scoring affinity of CRP for candidate *H. influenzae* CRP-binding sites, I used it to determine whether I_{seq} scores for *H. influenzae* CRE sequences significantly different from I_{seq} scores of characterized *E. coli* CRP-binding sites or candidate *H. influenzae* CRP-binding sites. As is evident in Fig. 1, scores for *H. influenzae* CRE sequences fall well within the "CRP site" score range. These observations imply that the *H. influenzae* cAMP–CRP complex will recognize, bind and activate transcription from these sequences. Non-parametric pairwise comparisons showed that sequences identified as CRE in *H. influenzae* competence gene promoters do not score significantly differently from the set of known *E. coli* or candidate *H. influenzae* CRP sites (Table 4).

Discussion

REGULATORY ELEMENTS (CRE) IN PROMOTER REGIONS OF COMPETENCE GENES ARE BINDING SITES FOR cAMP–CRP

H. influenzae and *E. coli* CRP-binding site sequences are expected to be very similar. The *H. influenzae* CRP amino acid sequence shows 87% similarity (77% identity) to its *E. coli* homologue, and the single amino acid difference in the DNA-binding helix-turn-helix domain is at a position predicted to be unimportant for DNA-binding (Wintjens & Rooman, 1996). Moreover, Chandler (1992) has demonstrated that the *E. coli crp* gene can complement an *H. influenzae crp* strain. The expected binding site sequence similarity was confirmed by the discovery of sequences resembling the *E. coli* CRP-binding site in the *H. influenzae* genome in the promoter regions of a number of genes whose transcription is expected to be activated by cAMP–CRP (Macfadyen *et al*., 1996) (Table 3). The presence of CRP-binding sites in the promoter region of a gene or genes required for competence was predicted, because cAMP and CRP are absolutely required for competence (Chandler, 1992; Dorocicz *et al*., 1993). I have now demonstrated that the cAMP–CRP complex is likely to bind candidate regulatory sites in the promoter region of several competence genes. The corollary—that these sequences are NOT primarily binding sites for the Sxy protein—is consistent with the observation that the predicted Sxy amino acid sequence possesses no identifiable helix-turn-helix DNA-binding motifs (as determined by a motif sequence search using the Emotif search algorithm (Nevill-Manning *et al*., 1999; L. Bannister, unpublished data). Moreover, the *H. influenzae* genome encodes only one other member of the CRP-like family (Henikoff & Henikoff, 1994) of helix-turn-helix DNA-binding proteins, FNR. Disruption of the *fnr* gene has no effect on competence development (C. Ma, unpublished data), implying that the FNR protein plays no role in transcriptional activation of competence genes.

In contrast with the simple regulatory cascade model proposed by others (Tomb *et al*., 1993; Karudapuram & Barcak, 1997) (in which transcription of Sxy is activated when cAMP levels rise, and Sxy subsequently activates transcription of other competence-specific genes), I suggest that the cAMP–CRP complex directly activates transcription of numerous competence-associated genes, including *sxy*, under conditions where intracellular cAMP levels rise. One or more of these competence-associated genes may encode a secondary regulator of competence development and allow fine-tuning of competence induction according to the nature of the cellular environment, as described below.

A COMPETENCE SPECIFIC REGULATOR MAY PROMOTE CRP BINDING TO CRE SEQUENCES

Experimental observations also support a model in which a secondary transcriptional activator regulates expression of competence genes, in addition to cAMP–CRP, under nutrient-limited or "starvation" conditions (Dorocicz *et al*., 1993). While 25–100% of *H. influenzae* cells become competent under nutrient-limited conditions ($TF_{max} = 10^{-2}$) (Goodgal & Herriott, 1961), only 1% of cells ($TF_{max} = 10^{-4}$) become competent at the onset of stationary phase growth in rich medium (R. J. Redfield, pers. comm.). Although CRP and adenylate cyclase activity are required for competence development under both sets of conditions, addition of cAMP to a stationary phase culture does not further boost competence, suggesting that the higher level of competence developed by nutrient-limited cells is not simply the result of higher endogenous levels of cAMP.

If a starvation-specific regulator of competence genes exists, it might be expected to act like other known transcriptional co-activators such as CytR (Pedersen & Valentin-Hansen, 1997) by independently binding cognate regulatory sequences in the promoters of regulated genes. However, the iterative algorithm "Consensus" (Stormo & Hartzell, 1989; Hertz *et al*., 1990; Hertz & Stormo, 1999) failed to identify any other conserved consensus sequence shared by the promoter regions of the competence genes listed in Table 1 (G. Stormo, pers. comm.). This suggests that a second regulator of competence does not act simply by recognizing a binding site in the promoter regions of competence genes and activating transcription.

However, a second regulator of competence might act cooperatively with CRP by providing additional DNA–protein contacts to stabilize the transcription initiation complex. The observation of two further uniquely conserved minor sequence elements within CRE sequences supports this model. As well as possessing a change in consensus, from TGTGA to TGCGA, CRE sequences have much more highly conserved central and flanking sequences than other CRP-binding sites. While most *E. coli* CRP sites are merely A/T rich on the outside of the consensus, the flanking region of CRE sequences are highly conserved as 5′-TTT-3′. In the centre region the TC at bp 9–10 is also highly conserved. A true contrast between "normal" CRP sites and CRE sites is therefore represented as

CRP: WWWTGTGANNNNNNTCACAWWW,

CRE: TTTTGCGATCNNGATCGCAAAA,

where W = A or T. Could these extra-conserved regions of CRE sequences be recognition sequences for another DNA-binding regulator which may stabilize the CRP–DNA complex? A known example of such a regulator is the cII transcriptional activator of phage □□, whose binding site spans the consensus binding site for a primary regulator of transcription: cII interacts with a 5′-TTGC-3′ consensus sequence on either side of the −35 region of the promoter (RNA polymerase binding site) (Place *et al*., 1984). In an analogous model, a secondary regulator of competence might bind cooperatively with CRP, interacting with the TTTxxxxxTC portion of each CRE half-sequence, and thereby increase affinity of the CRP protein for these sequences (G. Stormo, pers. comm.).

Alternatively, a second transcriptional activator might act synergistically with CRP at CRP/CRE sites to promote transcription of cAMP–CRP-regulated competence genes by inducing a conformational change in CRP that increases its affinity for CRE sequences (although this model does not directly address the existence of the described extra-conserved regions of CRE sequences). All of the CRE sequences identified in promoters of competence genes (except for that of *sxy*) contain a conserved C (rather than T) at base 6 and a conserved G (rather than A) at base 17 (Table 1). Both changes fall within the most highly conserved motifs of the CRP-binding site consensus—the symmetric sequences predicted to interact with the helix-turn-helix motif of each subunit of the CRP dimer (Botsford & Harman, 1992). Changing bases at these positions to the *E. coli* consensus bases T and A, respectively, increases the I_{seq} score for each sequence by 4.13 (or an average of 27%), implying that these consistent differences in CRE sequences reduce affinity of the CRP protein for these sites. Sequence specificity of DNA recognition by CRP might, however, be altered by interaction with a second regulator. Such "protein-induced fit" (Pedersen & Valentin-Hansen, 1997) could increase the affinity of cAMP–CRP for CRE sites in competence-specific promoters.

COULD SXY BE THE SECOND REGULATOR OF COMPETENCE?

Overexpression of the *sxy* gene from a plasmid greatly increases competence (Williams *et al*., 1994), suggesting that the Sxy protein mediates the response to a competence-specific signal under nutrient-limited conditions, and that expression, stability or degree of activation of this proposed regulator may be the limiting factor in competence development. However, recent studies have demonstrated that expression levels of a *sxy*::*lacZ* fusion in late exponential phase growth in rich medium are similar to those measured after transfer to nutrient-limited medium, even though the latter treatment induces 100-fold higher competence (L. Bannister, unpublished data). This implies that if Sxy is acting as a second regulator of competence, the limiting factor may be its degree of activation or stability, rather than its transcription. Activated Sxy may act cooperatively with CRP to allow maximal expression of competence genes, as described above. Alternatively, Sxy may activate expression of another as yet uncharacterized competence regulator.

In conclusion, while Sxy is clearly essential for development of competence by *H. influenzae*, the suggestion that CRE sequences in the promoter regions of competence genes are Sxy binding sites lacks any supporting evidence. Moreover, this

study emphasizes the central role of the cAMP–CRP complex in transcriptional activation of numerous competence-associated genes, and places "competence for DNA uptake" within the suite of *H. influenzae* responses to nutritional stress.

I am grateful to Dr Rosemary Redfield, in whose laboratory this analysis was carried out, for invaluable guidance and editorial help, and to Caixia Ma and Laura Bannister for sharing their unpublished observations. Thanks are also due to Dr Sally Otto and Dr Deborah Wilson who provided vital help with statistical analyses, and Dr Gary Stormo, who gave helpful feedback on application of the method. I was supported by a Studentship from the Canadian Cystic Fibrosis Foundation.

REFERENCES

Berg, O. G. & von Hippel, P. H. (1987). Selection of DNA binding sites by regulatory proteins. I. Statistical–mechanical theory and application to operators and promoters. *J. Mol. Biol.* **193,** 723–750.

Berg, O. G. & von Hippel, P. H. (1988). Selection of DNA binding sites by regulatory proteins. II. The binding specificity of cyclic AMP receptor protein to recognition sites. *J. Mol. Biol.* **200,** 709–723.

Blattner, F., Plunkett, G. R., Bloch, C. *et al.* (1997). The complete genome sequence of *Escherichia coli* K-12. *Science* **277,** 1453–1474.

Botsford, J. L. & Harman, J. G. (1992). Cyclic AMP in prokaryotes. *Microbiol. Rev.* **56,** 100–122.

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

Clifton, S. W., McCarthy, D. & Roe, B. A. (1994). Sequence of the *rec-2* locus of *Haemophilus influenzae*: homologies to *comE*-ORF3 of *Bacillus subtilis* and *msbA* of *Escherichia coli*. *Gene* **146,** 95–100.

de Combrugghe, B., Busby, S. & Buc, H. (1984). Cyclic AMP receptor protein: role in transcription activation. *Science* **224,** 831–838.

Death, A. & Ferenci, T. (1993). The importance of the binding-protein-dependent Mgl system to the transport of glucose in *Escherichia coli* growing on low sugar concentrations. *Res. Microbiol.* **144,** 529–537.

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

Dougherty, B. A. & Smith, H. O. (1999). Identification of *Haemophilus influenzae* Rd transformation genes using cassette mutagenesis. *Microbiology* **145,** 401–409.

Fleischmann, R. D., Adams, M. D., White, O. *et al.* (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269,** 496–512.

GenBank, http://www.ncbi.nlm.nih.gov/

Goodgal, S. H. & Herriott, R. M. (1961). Studies on transformation of *Haemophilus influenzae*: I. Competence. *J. Gen. Physiol.* **44,** 1201–1227.

Gwinn, M. L., Stellwagen, A. E, Craig, N. L. *et al.* (1997). *In vitro* Tn7 mutagenesis of *Haemophilus influenzae* Rd and characterization of the role of *atpA* in transformation. *J. Bacteriol.* **179,** 7315–7320.

Gwinn, M. L., Ramanathan, R., Smith, H. O. *et al.* (1998). A new transformation-deficient mutant of *Haemophilus influenzae* Rd with normal DNA uptake. *J. Bacteriol.* **180,** 746–748.

Henikoff, S. & Henikoff, J. G. (1994). Protein family classification based on searching a database of blocks. *Genomics* **19,** 97–107.

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

Hertz, G. Z., Hartzell, G. W. & Stormo, G. D. (1990). Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput. Appl. Biosci.* **6,** 81–92.

Hertz, G. Z. & Stormo, G. D. (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, in press.

Karudapuram, S. & 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.

Karudapuram, S., Zhao, X. & 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.

Larson, T. G. & Goodgal, S. H. (1991). Sequence and transcriptional regulation of *com101A*, a locus required for genetic transformation in *Haemophilus influenzae*. *J. Bacteriol.* **173,** 4683–4691.

Macfadyen, L. P. (1999). Regulation of intracellular cAMP levels and competence development in *Haemophilus influenzae* by a phosphoenolpyruvate:fructose phosphotransferase system. Ph.D. Thesis, Department of Zoology, University of British Columbia, Vancouver, British Columbia, Canada. p. 207.

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

Macfadyen, L. P., Ma, C. & 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.

Mulligan, M. E., Hawley, D. K., Entriken, R. *et al.* (1984). *Escherichia coli* promoter sequences predict *in vitro* RNA polymerase selectivity. *Nucl. Acids Res.* **12,** 789–800.

Nevill-Manning, C. G., Wu, T. D. & Brutlag, D. L. (1999). EMOTIF Database, http://dna.Stanford.EDU/3motif

Notley, L. & Ferenci, T. (1995). Differential expression of *mal* genes under cAMP and endogenous inducer control in nutrient-stressed *Escherichia coli*. *Mol. Microbiol.* **16,** 121–129.

Pedersen, H. & Valentin-Hansen, P. (1997). Protein-induced fit: the CRP activator protein changes

sequence-specific DNA recognition by the CytR repressor, a highly flexible LacI member. *EMBO J.* **16,** 2108–2118.

PETERKOFSKY, A. & GAZDAR, C. (1971). Glucose and the metabolism of adenosine 3′:5′-cyclic monophosphate. *Proc. Natl. Acad. Sci. U.S.A.* **63,** 2794–2798.

PIFER, M. L. & SMITH, H. O. (1985). Processing of donor DNA during *Haemophilus influenzae* transformation: analysis using a model plasmid system. *Proc. Natl. Acad. Sci. U.S.A.* **82,** 3731–3735.

PLACE, N., FIEN, K., MAHONEY, M. E. *et al.* (1984). Mutations that alter the DNA binding site for the bacteriophage lambda cII protein and affect the translation efficiency of the cII gene. *J. Mol. Biol.* **180,** 865–880.

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

ROBISON, K. & CHURCH, G. DPInteract: a database on DNA-protein interactions. http://arep.med.harvard.edu/dpinteract/

SETLOW, J. K., NOTANI, N. K., MCCARTHY, D. *et al.* (1981). Transformation of *Haemophilus influenzae* by plasmid RSF0885 containing a cloned segment of chromosomal deoxyribonucleic acid. *J. Bacteriol.* **148,** 804–811.

SETLOW, J. K., SPIKES, D. & GRIFFIN, K. (1988). Characterization of the *rec-1* gene of *Haemophilus influenzae* and behavior of the gene in *Escherichia coli*. *J. Bacteriol.* **170,** 3876–3881.

SOKAL, R. R. & ROHLF, F. J. (1969). *Biometry*. New York: W. H. Freeman and Co.

STORMO, G. D. (1988). Computer methods for analyzing sequence recognition of nucleic acids. *Annu. Rev. Biophys. Chem.* **17,** 241–263.

STORMO, G. D. & HARTZELL, G. W. I. (1989). Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl. Acad. Sci. U.S.A.* **86,** 1183–1187.

TIGR (1999). The Institute for Genome Research, http://www.tigr.org

TOMB, J.-F., EL HAJJ, H. & SMITH, H. O. (1991). Nucleotide sequence of a cluster of genes involved in the transformation of *Haemophilus influenzae* Rd. *Gene* **104,** 1–10.

TOMB, J.-F., AMITAI, H. & ABUCHAKRA, S. (1993). A competence-specific regulatory sequence of *Haemophilus influenzae* (Abstract). *Molecular Genetics of Bacteria and Phages*. Cold Spring Harbor Laboratory, NY: Cold Spring Harbor Laboratory Press.

WILLIAMS, P. M., BANNISTER, L. A. & REDFIELD, R. J. (1994). The *Haemophilus influenzae sxy-1* mutation is in a newly identified gene essential for competence. *J. Bacteriol.* **176,** 6789–6794.

WINTJENS, R. & ROOMAN, M. (1996). Structural classification of HTH DNA-binding domains and protein-DNA interaction modes. *J. Mol. Biol.* **262,** 294–313.

ZAR, J. H. (1996). *Biostatistical Analysis*. Upper Saddle River, NJ: Prentice-Hall.

ZULTY, J. J. & 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.

AUTHOR QUERY FORM

HARCOURT PUBLISHERS

JOURNAL TITLE: JTBI DATE:18/9/2000
ARTICLE NO. : 20002179

Queries and / or remarks

Manuscript Page/line	Details required	Author's response
15	Blattner et al. (1997) Please furnish names of all authors.	
16	Fleischmann et al. (1995) → Please provide names of other authors.	
16	Gwinn et al. (1997) & (1998), Please provide names of other authors.	
17	Hertz & Stormo (1999) → Please update.	
18	Mulligan et al. (1984) → Please provide names of all authors.	
18	Place et al. (1984), Please provide names of all authors.	
19	Setlow et al. (1988), Please provide names of other authors.	