Molecular Biology, Vol. 35, No. 6, 2001, pp. 853–861. Translated from Molekulyarnaya Biologiya, Vol. 35, No. 6, 2001, pp. 1001–1009.

UDC 577.218.577.577.121.9

Computer Analysis of Regulatory Signals in Bacterial Genomes. Fnr Binding Sites

A. V. Gerasimova, D. A. Rodionov, A. A. Mironov, and M. S. Gelfand

State Research Center for Biotechnology GosNIIGenetika, Moscow, 113545 Russia;
E-mail: misha@imb.imb.ac.ru

Received March 16, 2001

Abstract—Comparative approach to computer analysis of regulatory signals allows one to predict new signals in bacterial genomes with high accuracy. A prediction is reliable whenever candidate signals are consistently observed in several related genomes. We applied comparative approach to the analysis of the Fnr regulon of gamma-proteobacteria. Responding to changes in the aerobic/anaerobic state of the medium, the transcriptional factor Fnr regulates expression of many genes. We predicted Fnr binding sites in 12 genes regulated by Fnr, and identified 17 new operons as potential members of the Fnr regulon of *Escherichia coli*. In addition, we described the Fnr regulon of other gamma-proteobacteria.

Key words: Fnr, computer analysis, aerobic/anaerobic regulation, *Escherichia coli*

INTRODUCTION

The transcriptional factor Fnr regulates aerobic/anaerobic-dependent gene expression in gamma-proteobacteria. In *Escherichia coli*, expression of more than 120 Fnr-regulated genes depends on alternation of aerobic and anaerobic conditions [1, 2]. Fnr activates expression of several genes of anaerobic enzymes, such as nitrate and nitrite reductase (anaerobic respiration) and pyruvate-formate lyase (formate-acetyltransferase) (anaerobic fermentation). Moreover, Fnr can repress some genes of respiratory enzymes, such as cytochrome *d* oxidase and NADH dehydrogenase. The aim of this study was to analyze the Fnr regulon and find new potentially Fnr-regulated genes in the *Escherichia coli* genome and in the less studied genomes of other gamma-proteobacteria.

METHODS

We considered the genomes of seven related gamma-proteobacteria. Full nucleotide sequences of the *Escherichia coli* [3], *Haemophilus influenzae* [4], *Vibrio cholerae* [5], and *Pseudomonas aeruginosa* [6] genomes were extracted from GenBank (http://www.ncbi.nlm.nih.gov/GenBank/) [7]. Preliminary nucleotide sequences of *Salmonella typhi*, *Klebsiella pneumoniae*, and *Yersinia pestis* were extracted from http://www.sanger.ac.uk/Projects/Microbes/ and http://genome.wustl.edu/gsc/Projects/bacteria.shtml. Bacterial genomes were analyzed using the software package Genome Explorer [8]. The training set was compiled using information on the Fnr sites obtained from the database of regulatory sites DPInteract (http://arep.med.harvard.edu/dpinteract/) [9]. The recognition rule was constructed using the program SignalX [8]. Functional annotation of genomes was done using the program BLASTA (http://www.ncbi.nlm.nih.gov/BLAST/) [10] and the data bank of protein amino acid sequences SWISSPROT (http://expasy.hcuge.ch/sprot/) [11]. Alignments of intergene regions in related organisms were constructed using the program Menteric (http://globin.cse.psu.edu/enterix/menteric/menteric.html) [12].

RESULTS AND DISCUSSION

Basing on of the training set, we constructed a weight matrix using the program SignalX:

a	*c*	*g*	*t*
−0.09	−0.09	−0.32	0.49
−0.29	0.05	−0.29	0.52
−0.26	−0.26	0.55	−0.03
0.40	−0.15	−0.37	0.12
−0.39	0.07	−0.06	0.38
−0.08	−0.01	0.02	0.07
0.07	−0.26	0.00	0.19
0.19	0.00	−0.26	0.07
0.07	0.02	−0.01	−0.08
0.38	−0.06	0.07	−0.39
0.12	−0.37	−0.15	0.40
−0.03	0.55	−0.26	−0.26
0.52	−0.29	0.05	−0.29
0.49	−0.32	−0.09	−0.09

0026-8933/01/3506-0853$25.00 © 2001 MAIK “Nauka /Interperiodica”

Table 1. Known and potential Fnr-regulated operons in the genomes of *Escherichia coli* (*E. c.*), *Salmonella typhi* (*S. t.*), *Klebsiella pneumoniae* (*K. p.*), *Yersinia pestis* (*Y. p.*), *Haemophilus influenzae* (*H. i.*), *Vibrio cholerae* (*V. c.*), and *Pseudomonas aeruginosa* (*P. a.*)

E. c.	*S. t.*	*K. p.*	*Y. p.*	*V. c.*	*H. i.*	*P. a.*	Function	Metabolic pathway	Fnr influence on gene expression
				First group					
narXL*/*narKGJI	***narXL*/ /*narKGJI***	***narXL*/ /*narKGJI***	# # *narX*	# # ***narX***	#/#	***narX*/*narK*** ***narG***	Nitrate reductase, nitrate transporter	Anaerobic respiration	Activates
nirBD; *nirC*	***NirBD***	***nirBD***	***nirBD***	#	#	#	Nitrite reductase	Denitrification	Activates
cydAB	***cydAB***	***cydAB***	***cydAB***	***cydAB***	***cydAB***	#	Cytochrome *d* oxidase	Aerobic respiration	Represses
ndh	***Ndh***	***ndh***	***ndh***	***ndh***	*ndh*	*ndh*	NADH dehydrogenase	Aerobic respiration	Represses
nrfABC-DEFG	***nrfABC-DEFG***	#	#	#	***nrfABCD***	#	Formate-dependent nitrite reductase	Anaerobic respiration	Activates
fdnGHI	*FdnG*	***fdnG****?*	#, ***fdoGHI***	# #	*fdnGHI*	#, ***fdoGHI***	Formate dehydrogenase	Anaerobic respiration	Activates
focA-pflB	***FocA-pflB***	***focA-pflB***	***focA-pflB***	*focA-pflB*	***focA-pflB***	#	Formate transporter, pyruvate-formate lyase	Anaerobic fermentation	Activates
ansB	*AnsB*	***ansB***	*ansB*	#	*ansB*	***ansB***	L-Asparaginase	Catabolism of asparagine	Activates
				Second group					
pdhR-aceEF- *lpdA*	***pdhR-aceEF-*** *lpdA*	***pdhR-*** *aceEF-lpdA*	***pdhR-*** *aceEF-lpdA*	***pdhR-aceEF-*** *lpdA*	# *aceEF*, *lpdA*	*pdhR-aceEF-lpdA*	Pyruvate dehydrogenase	Pyruvate metabolism	Represses
feoAB	***FeoAB***	*feoAB*	***feoAB***	***feoB***	#	***feoAB***	Ferrous iron transporter		
nrdDG	***NrdDG***	***nrdD***, *nrdG*	***nrdDG***	***nrdDG***	*nrdD*, *nrdG*	#	Anaerobic ribonucleoside-triphosphate reductase	Nucleoside metabolism	Activates
dmsABC	*dmsABC*	***dmsA, dmsBC***	***dmsABC***	#	***dmsABC***	# #	Anaerobic dimethyl sulfoxide reductase	Anaerobic growth on various sulfoxides	Activates
dcuA	***dcuA***	***dcuA***	***dcuA***	*dcuA*	#	#	C4-dicarboxylate anaerobic transporter	The Krebs cycle (transport)	Activates
fnr	***fnr***	***fnr***	***fnr***	***fnr***	***fnr***	*fnr*	Fnr transcription regulator		Represses
arcA	***arcA***	***arcA***	***arcA***	***arcA***	***arcA***	**#**	ArcA transcription regulator		Activates
yfiD	***yfiD***	***yfiD***	***yfiD***	***yfiD***	***yfiD***	**#**	?		Represses
tdcABC	***tdcABC***	***tdcABC***	#	#	#	#	tdcABC operon transcriptional activator	Catabolism of threonine	Activates
dcuC	*dcuC*	***dcuC***	#	***dcuC***	#	#	C4-dicarboxylate anaerobic transporter	The Krebs cycle (transport)	Activates

Table 1. (Contd.)

E. c.	*S. t.*	*K. p.*	*Y. p.*	*V. c.*	*H. i.*	*P. a.*	Function	Metabolic pathway	Fnr influence on gene expression
Third group									
b2503	***b2503***	***b2503***	#	#	#	#	?		
ompW (yciD)	***yciD***	***yciD***	***yciD***	***yciD***	#	#	Outer membrane protein W (porin)		
pyrG	***pyrG***	***pyrG***	*pyrG*	***pyrG***	***pyrG***	*pyrG*	CTP synthetase	Nucleoside metabolism	
upp	***upp***	***upp***	***upp***	*upp*	*upp*	*upp*	Uracil phosphoribosyl-transferase	Nucleoside metabolism	
b0780/moaA	***b0780/moaA***	***b0780/moaA***	***b0780/moaA***	*b0780/moaA*	#/***moaA***	# #	Molybdenum cofactor of protein A biosynthesis		
ccpR (yhjA)	***yhjA***	***yhjA***	#	***yhjA***	#	***yhjA***	Cytochrome *c*551 peroxi-dase	Antiperoxide protection	
ppsA	***ppsA***	***ppsA***	***ppsA***	*ppsA*	#	*ppsA*	Phosphoenolpyruvate syn-thase	Gluconeogenesis	
b0873-b0872	***b0873***	***b0873***	***b0873***	#	#	#	?, NADH oxidoreductase		
fadL	***fadL***	***fadL***	***fadL***	***fadL***	***fadL***	#	Long-chain fatty acid trans-porter	Fatty acid metabolism	
wrbA	***wrbA***	***wrbA***	*wrbA*	*wrbA*	#	*wrbA*	Trp repressor binding pro-tein		
sfhB	***sfhB***	*sfhB*	***sfhB***	***sfhB***	***sfhB***	*sfhB*	?		
gltX	***gltX***	***gltX***	*gltX*	*gltX*	*gltX*	*gltX*	Glutamyl-tRNA synthetase		
mltA	*mltA*	*mltA*	*mltA*	***mltA***	***mltA***	*mltA*	Murein transglycosylase A		
mtlA	*mtlA*	***mtlA***	***mtlA***	***mtlA***	#	#	Mannitol-specific trans-porter	Sugar metabolism	
b1973	***b1973***	***b1973***	#	#	#	#	?		
yjiO	***yjiO***	***yjiO***	#	#	#	#	Antibiotic transporter	Drug resistance	
ycfC	*ycfC*	*ycfC*	***ycfC***	#	***ycfC***	*ycfC*	?		
last	***last***	***last***	#	#	#	#	?		

Note: Boldface means that the found Fnr signal is conserved in this genome; # means that there is no orthologous gene for this organism.

Table 2. Known and potential Fnr boxes in the genomes of *Escherichia coli* (*E. c.*), *Salmonella typhi* (*S. t.*), *Klebsiella pneumoniae* (*K. p.*), *Yersinia pestis* (*Y. p.*), *Haemophilus influenzae* (*H. i.*), *Vibrio cholerae* (*V. c.*), and *Pseudomonas aeruginosa* (*P. a.*)

Gene	Genome	Nucleotide sequence of the Fnr box	Fnr		CRP*	
			position	weight	position	weight
		First group				
narK	*E. c.*	TTGATTTAcATCAA	–74	5.15		–
	E. c.	aTGATaaAtATCAA	–112	4.20		
	S. t.	TTGATTTAtATCAA	–74	5.05	–3	3.62
	K. p.	TTGATaTAAATCAA	–74	5.05		–
	P. a.	TTGATTcctATCAA	–75	4.41	–79	**3.67**
narG	*E. c.*	TTGATcgttATCAA	–106	4.66		–
	S. t.	TTGATcgttATCAA	–106	4.66		–
	K. p.	TTGATcgctATCAA	–107	4.59		–
nirB	*E. c.*	TTGATTTAcATCAA	–73	5.15		–
	S. t.	TTGATTTAcATCAA	–70	5.15		–
	K. p.	TTGATTTAcATCAA	–70	5.15		–
	Y. p.	TTGATTTAcATCAA	–69	5.15	28	3.61
ndh	*E. c.*	TTGATTaAcATCAA	–151	5.03	–155	**3.83**
	S. t.	TTGATgcAcATCAA	–151	4.65	–155	**3.88**
	S. t.	TTGtTgTtAATtAA	–70	3.94		
	K. p.	TTGATgcAcATCAA	–149	4.65	–150	**3.88**
	Y. p.	TTGATaTAtATCAA	–146	4.9		**3.76**
	V. c.	TTGATaaAtATCAA	–200	4.78		**3.83**
cydA	*E. c.*	TTGATaTttATCAA	–346	4.78		–
	E. c.	TTGtTcTcgATCAA	–294	4.57		
	S. t.	TTGATTTtAATCAA	–347	5.08		**3.67**
	S. t.	TTGtccgtgATCAA	–295	4.14	–555	4.09
	K. p.	TTGATTTAtATCAA	–346	5.05		–
	K. p.	TTGATcaccgTCgA	–246	3.98		
	Y. p.	TTGtTcTAAATCAA	–315	4.84	–424	3.92
	Y. p.	TTGtcgTAgATCAA	–264	4.48	–387	3.92
	V. c.	TTGATTTAgATCAA	–221	5.12	–382	3.99
	V. c.	TTGATTTAgATCAt	–271	4.54		
	V. c.	TTGATTgtttTCAA	–337	3.97		
	H. i.	TTGATcTAAgTCAA	–293	4.81		**3.68**
nrfA	*E. c.*	TTGATTaAAgaCAA	–142	4.49		–
	S. t.	TTGATTaAAgaCAA	–144	4.49	–781	3.5
	H. i.	TTtATTTAAAaCAA	–117	4.44		**4.12**
	H. i.	TTGATcaAgcTCAA	–67	4.48	87	3.91
ansB	*E. c.*	TTGtTTaAcgTCAA	–70	4.44	–124	3.53
	K. p.	TTGATTaAtgTCtA	–45	3.81		–
	P. a.	TTGcTgggcATCAA	–927	3.91		–
focA	*E. c.*	aTGATcTAtATCAA	–73	4.39		**3.59**
	S. t.	aTGATcTAtATCAA	–70	4.39		**4.07**
	K. p.	cTGATgaAAgaCAA	–244	3.86		–
	Y. p.	aTGATccAtATCAA	–78	3.94		**3.81**
	H. i.	TTGtgaatAATCAA	–320	4.09	–437	3.82
	H. i.				–217	3.53
fdnG	*E. c.*	TTGAggTAggTCAA	–134	4.32		–
	K. p.	cTGATcgAAAaCAA	–200	4.07		–
	H. i.	aTGATcTAgATCAc	–210	3.65		**4.96**
		TTtAacgAAATCAA	–82	3.58		**3.45**

Table 2. (Contd.)

Gene	Genome	Nucleotide sequence of the Fnr box	Fnr		CRP*	
			position	weight	position	weight
		Second group				
yfiD	*E. c.*	TTGATTTAAATCAA	–120	5.20		**3.82**
	E. c.	TTGATgTAAAaCAA	–173	4.87		
	S. t.	TTGATTTAAATCAA	–112	5.20		**3.76**
	S. t.	TTGtTTTAcATCAA	–165	4.87		
	K. p.	TTGATaTAAATCAA	–107	5.05		–
	K. p.	TTGtTTTAcATCAA	–160	4.87		
	Y. p.	TTGATaTAAAaCAt	–128	4.72		**3.55**
	Y. p.	TTGATaTAAAaCAt	–181	4.19	–48	3.92
	V. c.	TTGATTTAggTCAA	–598	4.81		**3.85**
	V. c.	TTGATTTgtgTCAA	–438	4.29		**3.5**
	H. i.	TTaATTTAgATCAA	–143	4.31		**4.38**
	H. i.	TTcATTatAAaCAA	–230	3.87		
narX	*E. c.*	TTGATgTAAAaCAA	–278	5.15		–
	E. c.	TTGATaTttATCAt	–240	4.20		
	S. t.	TTGATaTAAATCAA	–278	5.05	–321	3.62
	K. p.	TTGATaTAAATCAA	–250	5.05		–
	V. c.	TTGtTTTggATCAA	–129	4.39		
	P. a.	TTGATaggAATCAA	–113	4.41		
nrdD	*E. c.*	TTGAgcTAcATCAA	–248	4.63		**4.22**
	S. t.	TTGtTcTAcATCAA	–247	4.79		–
	K. p.	TTGtTcTgggTCAA		4		–
	Y. p.	TTGtTcTAggTCAA	–193	4.45		–
	V. c.	TTGATcTAAATCAA	–126	5.12		–
	H. i.	TTGATaTtAATCAg	174	4.35	–207	4.67
					–153	3.71
arcA	*E. c.*	TTGATaTAtgTCAA	–290	4.59		–
	S. t.	TTGATaTAtgTCAA	–289	4.59		–
	K. p.	TTGATaTAtgTCAA	–289	4.59	–558	3.54
	Y. p.	TTGATaTAtgTCAA	–296	4.59		**3.6**
	V. c.	TTGATgTAAATCAA	–254	5.15		–
	H. i.	TTGtTTTttATCgA	–264	4.18	–129	4.66
b0621 (*dcuC*)	*E. c.*	TTGATTTtAATCAg	–102	4.50		–
	S. t.	gTGATTTtAATCAg	–99	3.69	–361	3.63
	K. p.	TTGATTTgcATCAg	–105	4.21		–
	V. c.	TTGcTTTAgATCAt	–102	3.99		–
fnr	*E. c.*	TTGAcaaAtATCAA	–32	4.47		–
	S. t.	TTGAcaaAtATCAA	–31	4.47		–
	K. p.	TTGAccaAtATCAA	–29	4.54		–
	K. p.	cTGtTTTtAATCAA	–131	4.22		
	Y. p.	TTGAcgcAtATCAA	–33	4.24		**3.51**
	V. c.	TTGAcgTAcATCAA	–33	4.79		–
	H. i.	TTGcgTTAgATCAA	–40	4.13		**3.81**
pdhR	*E. c.*	aTGATTTcAATCAA	–116	4.43	–132	3.88
	S. t.	cTGATTTcAATCAA	–117	4.43		–
	K. p.	cTGATTTcAATCAA	–133	4.43		–
	Y. p.	aTGATTTcggTCAA	–133	4.04	–50	3.77
	V. c.	aTGATTTAggTCAA	–121	4.23		3.76
aceE	*E. c.*	aTGtTgTAAATCAA	–132	4.29		–
	V. c.	aTGATTTAggTCAA	–941	4.23		3.76
dcuA	*E. c.*	TTGtTaaAAAaCAA	–24	4.37		–
	S. t.	TTGtTaaAcAaCAA	–23	4.32	–2	3.62
	K. p.	TTGtTaaAAAaCAA	–22	4.37		–
	Y. p.	TTGAaTgAAATCAA	–405	4.24		–

Table 2. (Contd.)

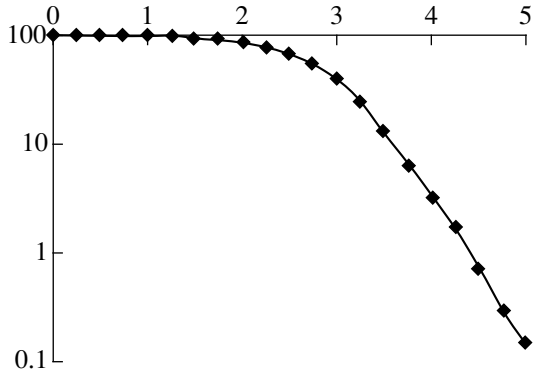
Gene	Genome	Nucleotide sequence of the Fnr box	Fnr		CRP*	
			position	weight	position	weight
feoA	*E. c.*	TTGAgaccgAaCAA	–204	4.18		**3.59**
	S. t.	TTGAgccAtATCAA	–214	4.08	–413	3.85
	K. p.	TTGcagcAtATCAA	–204	3.98		
	Y. p.	TTGATaTccAatAA	–488	3.72		–
	P. a.		–134	3.46		–
dmsA	*E. c.*	TTGATaccgAaCAA	–272	4.05		–
	S. t.	TcGATaTAtATCAg	–154	3.85		**3.63**
	K. p.	aTGATaatcATCgA	–589	3.71		–
	Y. p.	TTGATTccAgaCAA	–308	3.97		–
	H. i.	TTGATTTggcTCAA	–87	4.23		**3.68**
tdcA	*E. c.*	TTGAcaaAAATCAg	–182	4.04	–81	4.2
	S. t.	TTGATTgAAATCAg	–182	4.43	–81	4.2
	K. p.	TTGATTTtAATCAA	–521	5.08	–80	3.93
	K. p.				–52	3.61
		Third group				
b2503	*E. c.*	TTGATaTAtATCAA	–140	4.90		–
	S. t.	TTGAcTTAAATCAA	–141	4.89	–91	3.75
	K. p.	TTGATTactATCAA	–134	4.74		–
upp	*E. c.*	TTGAcTaAAgTCAA	–67	4.46		–
	S. t.	TTGATccAggTCAA	–74	4.28		**3.51**
	K. p.	TTGATaTAcgTCAA	–73	4.69	–45	3.65
	Y. p.	TTGATcTgAATCAg	–69	4.09		–
b0780	*E. c.*	TTGATaTAcATCAt	–146	4.42		–
	S. t.	TTGATaTAtATCAt	–144	4.32		–
	K. p.	TTGATgTcggTCAg	–92	3.99		–
	Y. p.	TTGtTTaAAATCAc	–188	3.99	–195	3.51
moaA	*E. c.*	aTGATgTAtATCAA	–264	4.42		–
	S. t.	aTGATaTAtATCAA	–265	4.32		–
	K. p.	cTGAccgAcATCAA	–255	3.99		
	Y. p.	gTGATTTtAAaCAA	–356	3.99	–550	3.65
	H. i.	aTGATTTAAATCAA	–343	4.62		**3.98**
	H. i.	cTGATTTtcATCAA	–195	4.45		
yciD	*E. c.*	TTGATTTAAATCAc	–163	4.39		**4.56**
	S. t.	TTaATcTggATCAA	–118	3.78	–83	3.57
	K. p.	TTGATTTcAcTCAt	–284	3.99	–138	3.64
	k. p.	TTaATcTggATCAA	–117	3.78	–82	3.56
	Y. p.	aTGATccAgATCAA	–124	4.01		**3.92**
	V. c.	TTGATTTccATCAA	–91	4.96	–144	3.75
pyrG	*E. c.*	TTGATTTgcgTCAA	–175	4.39		–
	S. t.	TTGATTTAcgTCAA	–174	4.84		–
	K. p.	cTGATTTAcgTCAA	–230	4.26		–
	V. c.	TTGATTTgAAgCAA	–134	4.2		**3.6**
	H. i.	TTGAcTTAgATCAA	–372	4.81	–440	4.79
	H. i.				–376	3.68
fadL	*E. c.*	aTGATcTAAAaCAA	–238	4.26		**3.76**
					–264	4.42
	S. t.	aTGATcTAAAaCAA	–237	4.26		**3.69**
					–263	4.36
	K. p.	TTGATTTAggaaAA	–78	3.97		–
	Y. p.	TTttTTgAgATCAA	–177	4.07		–
	V. c.	TTGATcTtgATgAA	225	4.11		–
	H. i.	TTtATTTAtAaCAA	–30	4.19	–185	3.64

Table 2. (Contd.)

Gene	Genome	Nucleotide sequence of the Fnr box	Fnr		CRP*	
			position	weight	position	weight
ppsA	*E. c.*	TcGATgTccAaCAA	–2	4.16		–
	S. t.	TcGATgTccAaCAA	–1	4.16		–
	K. p.	TcGATgTccAaCAA	–1	4.16		–
	Y. p.	TcGATgTccAaCAA	47	4.16		–
b0873	*E. c.*	TTGcgcTAAATCAA	–103	4.13		–
	S. t.	TTGcgcTAAATCAA	–102	4.13		–
	K. p.	TTGcgcTAAATCAA	–100	4.13		–
	Y. p.	TTGcgTcAAATCAA	–113	3.76		–
wrbA	*E. c.*	TTGtTaTAAATCAA	–114	4.77		**3.65**
	S. t.	TTGtTaTAAATCAA	–110	4.77		–
	K. p.	TTGtTaTAAATCAA	–112	4.77		–
sfhB	*E. c.*	TTGAcTTtccTCAA	–127	4.28		–
	S. t.	TTGAcTTtccTCAA	–123	4.28		–
	Y. p.	TTGATTatccTCgA	–127	4.00		
	V. c.	TTGAaTTAAcTCAA	15	3.99		
	H. i.	TTGtTcTtgATaAA	–313	4.06		
gltX	*E. c.*	TTcATgaAAATCAA	–2	4.22		–
	S. t.	TTcATgaAAATCAA	–1	4.22		–
	K. p.	TTcATgaAAATCAA	–4	4.22		–
b1973	*E. c.*	TcGtTTgtcATCAA	–138	4.09		–
	S. t.	TTGATaTcAAaaAA	–131	4	–26	3.68
	K. p.	TTGATTgAtATCgt	–166	3.81		–
yjiO	*E. c.*	TTGATTaAccgCAA	–281	4.04		–
	S. t.	TTGATTaAcATCAA	–289	5.03		**3.85**
	K. p.	TTGATTTttAaCAt	554	4.07		–
ycfC	*E. c.*	TTtAcTTAAAaCAA	–26	4.03		–
	Y. p.	TTGATggAAATaAA	–589	4.38		–
	Y. p.	cTGATccAAgTCAA	–67	3.78		
	H. i.	TTtATTTggATCAA	–357	4.09		
last	*E. c.*	TTGAcaTAtATCAA	–359	4.59		–
	S. t.	TTGAcaTAtATCAA	–359	4.59		–
	K. p.	TTGAcaTAtATCAA	–356	4.59	–95	3.54
mtlA	*E. c.*	TTGATaTcAcaCAA	–155	4.14		**4.74**
	S. t.			–	–165	4.67
	K. p.	gTGATcTtAATCAA	–328	3.72		
	Y. p.	gTGATaaAtATCAA	–220	4.19		
	V. c.	aTGATTTtgAaCAA	–287	3.97		
mltA	*E. c.*	TcGcTaTtAATCAA	–136	4.14	–163	3.69
	V. c.	TTGATggAttTCAA	–33	3.91		
	H. i.	TTGATggAttTCAA	87	4.04		

Note: The cases of coincidence between the CRP and Fnr boxes are marked in bold.

* The search threshold for CRP boxes is equal to 3.5.



Weight distribution function for Fnr sites. Horizontal axis, the weight threshold value; vertical axis, percentage of genes selected at the given threshold, on a logarithmic scale.

The weight of each of the four nucleotides is given at each position of the matrix. The weights were defined by the following formula

$$W(b, k) = 0.25 \sum_{i = A, C, G, T} \log[(N(b, k) + 0.5)/(N(i, k) + 0.5)],$$

where $N(b, k)$ is the count of nucleotide b at position k. The average weight W on the Bernoulli random sequence equals 0. The base of the logarithm is chosen such that the variance equals 1. In other words, W is the Z-statistics, and the probability of random occurrence of a signal can be assessed using the Gaussian distribution.

The score of a candidate site is the sum of positional weights of the constituent nucleotides:

$$S(b_1, \ldots, b_n) = \sum_{k = 1}^{n} W(b_k, k).$$

In these terms, the signal (Fnr box) is palindromic. Hence, the matrix obtained improves the already known consensus sequence TTGATnnnnATCAA [13]. Then, we searched upstream gene regions for potential Fnr boxes. As a result, candidate Fnr boxes were found upstream of 121 genes when the threshold was equal to 4.0. This choice of a threshold leads to the loss of some Fnr-regulated genes. We also searched for Fnr boxes using other values of the threshold, but the first choice appeared to be optimal. When the value of the threshold exceeds 4.0, many known sites are lost (underprediction), whereas when it is less than 4.0, almost all genes have potential Fnr boxes (overprediction). Indeed, consider the probability distribution of scores in the *E. coli* genome (figure). One can see that when the threshold equals 4.0, about 5% of all genes in the genome are selected. At the same time, it appears to be impossible to formally assess the type 1 and type 2 errors, because we do not know exactly what genes compose the regulon. When constructing the weight matrix, we used sites from 9 out of 121 selected genes (obtained from the database of regulatory sites DPInteract). The fact that 12 genes are Fnr-regulated was experimentally confirmed. We applied the standard procedure of comparison between related genomes [14] and found out that the genomes of the considered bacteria (*S. typhi, K. pneumoniae, Y. pestis, H. influenzae, V. cholerae*, and *P. aeruginosa*) contain genes orthologous to *fnr*. This suggests that the Fnr regulon of these gamma-proteobacteria is conserved. Orthologs of 121 genes of *E. coli* with potential Fnr boxes upstream were identified in the genomes of other bacteria. We analyzed the upstream regions of the orthologous genes using the weight matrix. The 39 genes having upstream regions with potential Fnr boxes retained in at least three considered genomes (one of them is *E. coli*) are listed in Tables 1 and 2. These genes were divided into three groups. The first group includes 9 genes. Their sites were used to construct the recognition matrix. The second group consists of 12 genes that were experimentally demonstrated to be under regulation by Fnr but the corresponding Fnr boxes were not identified. The third group consists of 18 genes. They have upstream regions with potential Fnr boxes retained in at least two considered genomes in addition to *E. coli*.

In order to estimate the number of false positives, we used a statistical model. The estimates were as followings: $P = 5\%$ is the fraction of genes with candidate sites (figure), $g = 2/3$ is the fraction of orthologs in any two genomes, $n = 4000$ is the average number of genes in a genome. The expected number of false positives can be expressed as follows:

$$C_n^k n g^2 P^3 = (5 \times 6)/(1 \times 2) \times 4000 \times 4/9 \times 125 \times 10^{-6} = 3.3.$$

It is known that the Fnr protein is homologous to the regulator CRP, and their signals resemble one another [2]. Thus, it is possible that a number of the predicted sites are CRP binding sites. Using the weight matrix for CRP boxes [15], we obtained scores of the candidate CRP binding sites of genes from the Fnr regulon. Many known genes from the Fnr regulon (the first and the second groups of Table 2) have potential CRP boxes. In addition, it is known that the genes *ansB* and *tdcA* are under regulation by both CRP and Fnr [13, 16]. CRP also regulates the *mtlA* gene that belongs to the third group [17]. We observed numerous candidate CRP boxes upstream of this gene, and their scores were higher than those of candidate Fnr boxes. The hypothetical *b2503* gene was predicted to be the Fnr-regulated, because the alignment of the upstream regions of the orthologous genes revealed some conservatism in the area of the potential Fnr box.

Apparently, the *aldA* gene also belongs to the Fnr regulon. This gene does not have orthologs in the considered genomes and thus, using the formal approach, cannot be assigned to that regulon. However, there is a potential Fnr box upstream of *aldA* in the *E. coli* genome. In addition, this gene is known to be under regulation of ArcA and CRP [18] (the CRP binding site does not coincide with the predicted Fnr box). The ArcA enzyme controls aerobic respiration and regulates expression of several genes of the Fnr regulon, the *arcA* gene itself is regulated by Fnr [2] (Table 2). Detailed analysis of the ArcA regulon is the subject of another paper.

Thus, we have identified Fnr binding sites in upstream regions of 12 *E. coli* genes regulated by Fnr, and found 17 additional genes that may belong to the Fnr regulon. We have described the Fnr regulons of *S. typhi, K. pneumoniae, Y. pestis, H. influenzae, V. cholerae,* and *P. aeruginosa*. Currently we analyze the regulatory system Anr (the ortholog of Fnr) in the genomes of pseudomonads.

ACKNOWLEDGMENTS

This work was partially supported by the Russian Foundation for Basic Research (projects nos. 99-04-48247 and 00-15-99362), INTAS (99-1476), and the Howard Hughes Medical Institute (55000309).

REFERENCES

1. Bauer, C.E., Elsen, S., and Bird, T.H, *Annu. Rev. Microbiol.*, 1999, vol. 53, pp. 495–523.
2. Lynch, A.S. and Lin, E.C.C., *Escherichia coli and Salmonella. Cellular and Molecular Biology*, Neindhard, F.C., Ed., Washington DC: ASM Press, 1996, pp. 1526–1538.
3. Blattner, F.R., Plunkett, G., Bloch, C.A., *et al.*, *Science*, 1997, vol. 77, pp. 1453–1474.
4. Fleischmann, R.D., Adams, M.D., White, O., *et al.*, *Science*, 1995, vol. 269, pp. 496–512.
5. Heidelberg, J.F., Eisen, J.A., Nelson, W.C., *et al.*, *Nature*, 2000, vol. 406, pp. 477–483.
6. Stolver, C.K., Pham, X.Q., Erwin, A.L., *et al.*, *Nature*, 2000, vol. 406, pp. 959–964.
7. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., *et al.*, *Nucleic Acids Res.*, 2000, vol. 28, pp. 15–18.
8. Mironov, A.A., Vinokurova, N.P., and Gelfand, M.S., *Mol. Biol.*, 2000, vol. 34, pp. 222–231.
9. Robison, K. and Church, G., *J. Mol. Biol.*, 1998, vol. 284, pp. 241–254.
10. Altschul, S.F., Madden, T.L., Schaffer, A.A., *et al.*, *Nucleic Acids Res.*, 1997, vol. 25, pp. 3389–3402.
11. Bairoch, A. and Apweiler, R., *Nucleic Acids Res.*, 2000, vol. 28, pp. 45–48.
12. Florea, L., Riemer, C., Schwartz, S., *et al.*, *Nucleic Acids Res.*, 2000, vol. 28, pp. 3486–3496.
13. Green, J., Irvine, A.S., Meng, W., and Guest, J.R., *Mol. Microbiol.*, 1996, vol. 19, pp. 125–137.
14. Mironov, A.A. and Gelfand, M.S., *Mol. Biol.*, 1999, vol. 33, pp. 127–132.
15. Gelfand, M.S., Novichkov, P.S., Novichkova, E.S., and Mironov, A.A., *Briefings in Bioinformatics*, 2000, vol. 1, pp. 357–371.
16. Chattopadhyay, S., Wu, Y., and Datta, P., *J. Bacteriol.*, 1997, vol. 179, pp. 4868–4873.
17. Ramseier, T.M. and Saier, M.H., *Microbiology*, 1995, vol. 141, pp. 1901–1907.
18. Pellicer, M.T., Lynch, A.S., De Wulf, P., *et al.*, *Mol. Gen. Genet.*, 1999, vol. 261, pp. 170–176.