Biophysics, Vol. 48, Suppl. 1, 2003, pp. S21–S25.

ArcA regulator of Gamma-Proteobacteria: Identification of the Binding Signal and Description of the Regulon

A. V. Gerasimova [1], M. S. Gelfand[1,2,3], V. Yu. Makeev[1], A. A. Mironov[1,3], and A. V. Favorov[1]

[1]State Scientific Center GosNIIGenetika, 1 1st Dorozhny proezd, Moscow, 117545 Russia; E-mail: a_gerasimova@yahoo.com
[2]Institute for Problems of Information Transmission, Russian Academy of Sciences, 19 Bolshoi Karetny per., Moscow, 127994 Russia
[3]Department of Bioengineering and Bioinformatics, Moscow State University, 1-73 Vorobievy Gory, Moscow, 119992 Russia

Received November 3, 2003; in final form, November 19, 2003

Abstract—Basing on a common signal previously identified in nine regions upstream of *Escherichia coli* genes regulated by ArcA, orthologous ArcA-regulated genes were identified in genomes of gamma-proteobacteria: *Escherichia coli*, *Yersinia pestis, Pasteurella multocida*, and *Vibrio vulnificus.* In addition to 10 genes in the training set, 23 new genes with conserved candidate ArcA-binding sites were identified. Fourteen of them have been described in the literature as transcriptionally regulated depending on the oxygen level. A candidate site upstream of *arcA* makes it plausible that this gene is autoregulated by the negative feedback loop mechanism.

Key words: ArcA, SeSiMCMC, protein-DNA binding, regulation of transcription, protein binding site, aerobic-anaerobic switch, Gibbs sampler, gamma-proteobacteria

INTRODUCTION

Depending on the oxygen concentration in the cytoplasm, metabolism of many bacteria, e.g., *Escherichia coli*, can be aerobic or anaerobic. The switch between these two states is mediated by several global systems of transcriptional regulation [1]. One such system is the two-component system ArcAB, where the membrane protein ArcB is a sensor, whereas ArcA is the DNA-binding regulator of transcription. In anaerobic conditions, ArcB autophosphorylates itself and then transphosphorylates ArcA, thus stimulating ArcA binding to the DNA. In aerobic conditions, ArcB does not autophosphorylate. Another system consists of one transcriptional regulator FNA. It is known that FNR regulates transcription of the gene *arcA*. Some genes are regulated by both FNR and ArcA [2]. Binding signals of FNR and ArcA have been described [3–5]. The FNR signal is a palindrome [3], whereas the ArcA signal is a direct repeat [5].

The aim of this study was to describe the ArcA regulon in the genomes of *E. coli, Yersinia pestis, Pasteurella multocida*, and *Vibrio vulnificus*, using the common recognition signal constructed previously [5]. These genomes represent three best -studied families of gamma-proteobacteria.

DATA AND METHODS

To determine the binding signal of ArcA, a training set of nine upstream regions of ten ArcA-regulated genes was selected from the DPInteract database (http://arep.med.harvard.edu/dpinteract/) [4]; one region was common to two genes forming a divergon (table, a). The SeSiMCMC program (http://bioinform.genetika.ru/SeSiMCMC) was used to find a common signal of length not exceeding 21 nt (two helix turns, that is, the upper limit of length for non-combined transcriptional signals in prokaryotes). Of the three possible types of symmetry

Table. ArcA –regulated genes in the genomes of *Escherichia coli*, *Yersinia pestis*, *Pasteurela multocida* and *Vibrio vulnificus* and positions of ArcA-boxes in the upstream regions; (a) genes of the *E. coli* training set and their orthologs; (b) candidate ArcA-regulated genes

Gene	Site position in *E. coli*	Site score in *E. coli*	*E. coli*	*Y. pestis*	*P. multocida*	*V. vulnificus*	References
(a) Training set							
cydA	–586	5.94 4.44	GTTAACTAAATGTTA TAACAaaataaTAAC	–345 => 4.60	–334 <= 4.40	–138 => 5.15	[15]
lldP	–134	5.94	TAACATTTAGTTAAC	*	–362 => 4.30 –139 => 5.03	–	[15]
glcD	–160	5.76	TAACATTgAGTTAAC	*	*	*	[16]
glcC	–105	5.76	GTTAACTcAATGTTA	*	*	*	[16]
sodA	–79	5.50	TAtCATTTAaTTAAC	–84 => 5.50	–63 <= 4.53	–	[17]
aldA	–49	5.08	TAACAaTgtaTTcAC	–209 <= 5.26 –137 <= 4.27	*	–117 => 5.50	[18]
lpda	–230 –232	4.42 5.06	TtAaAaaTtGTTAAC GTTtAaaAAtTGTTA	–136 => 4.68	–	–	[19]
icdA	–111	4.91	TtACAaaTcaTTAAC	–40 => 4.73	–133 => 4.42	–	[20]
gltA	–654 –393	5.15 4.34	GTTAACTttcTGTTAc AACATTTAtTTAAt	–422 => 4.40 –248 => 4.83 –248 => 4.83	–167 => 4.91 –112 => 5.07	–337 => 5.88 –231 => 4.04 –146 => 4.34	[15]
sdhC	–69 –330	5.15 4.34	TAACAgaaAGTTAACa TTAAaTAAATGTTg	–476 <= 4.83 –302 <= 4.40 –302 <= 4.40	*	–117 => 4.20	[15]
(b) Genes with conserved ArcA boxes							
cyoA	–249	4.95 4.92	GTTAAaTAAtTGTTt GaTAAtTAttTGTTA	–303 => 4.71	*	–147 => 4.48	[21]
fadD	–74 –72	5.17 4.79	GTTAAtattATGTTA TAAtATTatGTTAAC	–75 => 4.19 –77 <= 5.14	–	–133 => 4.53 –135 <= 4.71	[22]
focA	–219	4.28	tTTAAtTAAcTGTTt	–408 => 4.09 –284 <= 4.34	–202 => 4.09 –191 => 4.22 –80 => 4.16	91 <= 4.77	[23]
adhE	–304 –299	4.59 4.57	ctACAaTTtaTTAAC aTTtAtTAAcTGTTA	–395 <= 4.21	–	–261 => 4.58 –241 => 4.59	[24]
pdhR	–84	4.85	GTgAAtacttTGTTA	–102 <= 4.85	*	–197 <= 4.89	[25]
cstC	–38 –49	4.98 4,42	TAACATaTAaaTAAC TtACtTaTtaTTAAC	–177 => 4.21	*	–329 => 4.19	
mdh	–263	5.19	GTaAAtTAAtTGTTA	–71 => 4.22 –92 <= 4.22	–	–105 <= 4.88	[26]
argR	–186	5.19	TAACAaTTAaTTtAC	–386 => 4.22	–	–282 => 4.88	[27]
arcA	–565	4,67	TtACAaaTtcTTAAC	–699 <= 4.01	–	–353 <= 4.57	[15]
putP	–312	4.53	aTgAAtgAAATGTTA	–724 <= 4.14 –160 <= 4.16	–72 <= 4.20 –61 <= 4.67	–	[28]
prpR	–225	4.42	GTTAtCaActTGTTA	–233 => 4.14 –444 <= 4.25	*	–86 <= 4.24	

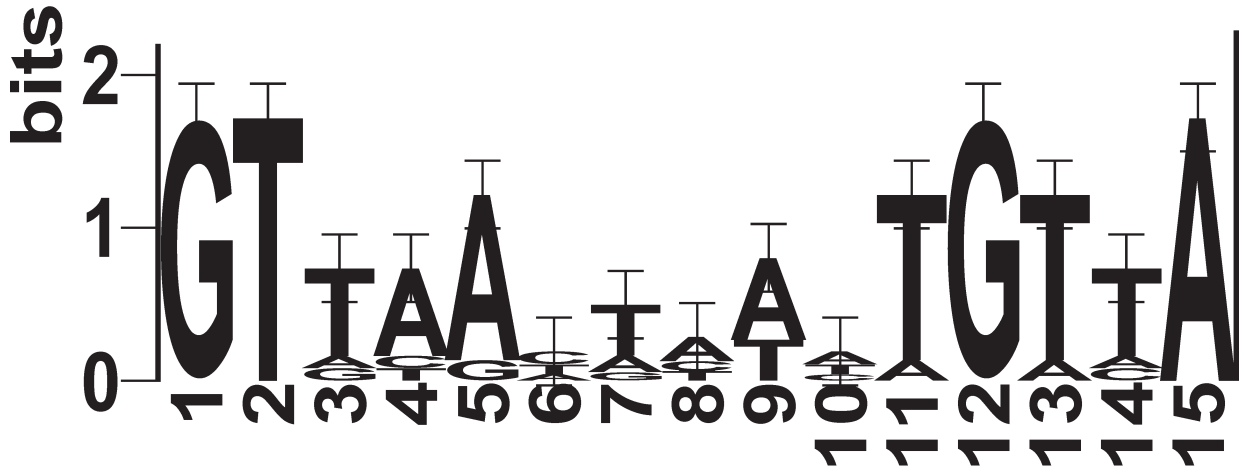
Table. (Contd.)

trxC	–65	4.42	TAACATaTtagaAAC	–24 => 4.04	–699 => 4.05 –404 <= 4.12	–486 => 4.21 –328 <= 4.40	[29]
ppa	–203	4.40	GTaAAtaAAAcGTTA	–116 <= 4.00	–150 => 4.27 –119 => 4.94	–	
tolB	–54	4.34	GTTAACattcTGcTA	–53 <= 4.08	–	–62 => 4.10	
slyB	–6	4.30	TttCAaTgAtTaAAC	–210 <= 4.34	–446 => 4.02	*	
glpF	–88	4.28	TAACgaTaAGTTtAC	–85 => 4.22	–	–289 <= 5.50	[30]
acpP	–62	4.27	cAACATTTtaTacAC	–62 => 4.27	–142 => 4.09	–	
phnA	–193	4.27	TAACATTatcTTAAa	–55 => 4.60	–99 => 4.24	–	
fadB	–77	4.26	TtgCATaTttTTAAC	–81 => 4.57	*	–158 <= 4.07	[22]
ompC	–508	4.96	GTTAAtTAttTGTgA	–152 <= 4.05	–238 => 4.98 –259 <= 4.00	*	[31]
dcuB	–390	4.95	GTTAAtTAAcTaTTA	–8 <= 4.34	–647 <= 4.16 –584 <= 4.16	–267 => 4.59 –130 => 4.07 –405 <= 4.10	[32]
pldA	–145	4.91	GTTAAtgAAATGTTg	–149 <= 5.15	–	–192 => 4.15	
ndk	–109	4.83	tTTAAaaAAATGTTA	–	20 => 4.11	–	

Notes: The star '*' denotes the absence of an ortholog, minus, '–' the absence of a ArcA-box, satisfying the recognition threshold (see Data and Methods). Signs '<=' and '=>' denote the direct and complementary strand respectively (relative to the direction of transcription of the regulated gene). The references are to the papers showing that the gene is regulated by the oxygen level in at least one of the four genomes.

considered—direct repeat, palindrome, non-symmetrical signal—the best results were obtained for the former [5], as shown in the figure. SignalX [6] was used to construct the recognition profile as described in [7]. The scanning of both chains of the *E. coli* [8] genome with this profile and threshold 4.25 yielded candidate ArcA-binding sites (ArcA-boxes) upstream of 257 genes.

Orthologs of *acrA* were identified in the genomes of *Y. pestis* [9], *P. multocida* [10], and *V. vulnificus* [11] downloaded from GenBank (http://www.ncbi.nlm.nih.gov/GenBank/) [12] using the standard procedure for identification of orthologous genes [13]. This allowed us to assume that the ArcA regulon is conserved in these species, and to apply comparative filtering of the results. At that, the Genome Explorer-



The ArcA-binding signal according to [5]. Horizontal axis: position. The total height of the symbols reflects the information content of the signal position; the height of a symbol is proportional to its positional frequency (the WebLogo program, http://www.bio.cam.ac.uk/cgi-bin/seqlogo/logo.cgi).

program [6] was used to analyze the complete genomes of *Y. pestis, P. multocida*, and *V. vulnificus* with the aim to find conserved members of the ArcA regulon. For each *E. coli* gene having an ArcA-box in the upstream region, orthologs were identified in the remaining three genomes. If at least two orthologs had a candidate ArcA-box at the threshold 4.0, the gene was assumed to have a conserved ArcA-binding site.

RESULTS

In the genome of *E. coli,* 23 genes with conserved ArcA sites were identified (table, b). Of these, 14 have been described in the literature as regulated by oxygen-dependent mechanisms. The training set consisted of ArcA-regulated genes, Known recognition signals of other oxygen-dependent regulators differ from the identified signal (figure). Further, this signal was evolutionarily conserved in several genomes. Thus, the signal indeed is recognized by the ArcA regulator.

Interestingly, one of the candidate members of the regulon was *arcA* itself. This allows us to suggest the mechanism of autoregulation of this gene by a negative feedback loop.

DISCUSSION

We were looking for genes having conserved upstream ArcA-boxes [5], and found 23 such genes. This set is enriched in genes whose expression is known to be oxygen-dependent. Indeed, the *E. coli* genome contains 4404 genes. Of the identified 23 genes, 14 belong to the oxygen-dependent set. Assuming a very liberal estimate that this set contains at most 500 genes, we obtain a two-by-two contingency table '14 9 // 500 3904'. Applying the Fischer criterion to this table, we reject the null hypothesis about independence of oxygen regulation and a conserved ArcA site at the significance level $2 \cdot 10^{-7}$. Making more plausible assumptions about the number of oxygen-dependent genes makes the result even more reliable. Nine genes not mentioned in the literature may thus be considered as new candidate members of the ArcA regulon.

One such gene is *arcA* itself. The candidate ArcA-box of *E. coli*, conserved also in *Y. pestis* and *V. vulnificus*, lies at a large distance from the gene start. It is well known that ArcA binding sites are often accompanied by sites binding IHF, an auxiliary transcription factor bending the double helix of DNA [33]. Thus the distance to the gene is not a crucial parameter for the signal. Moreover, it has been shown in experiment that *arcA* is regulated by FNR, which also binds to a distant site [14]. This makes the hypothesis of autoregulation of *ArcA* quite plausible.

ACKNOWLEDGMENTS

This study was partially supported by grants from HHMI (55000309), LICR (CRDF RB0-1268), and RFBR (02-04-49111).

REFERENCES

1. Iuchi, S. and Weiner, L., *J. Biochem.* (Tokyo), 1996, vol. 120, no. 6, pp. 1055–1063.
2. Unden, G., Becker, S., Bongaerts, J., *et al*., *Arch. Microbiol*., 1995, vol. 164, pp. 81–90.
3. Green, J., Irvine, A.S., Meng, W., and Guest, J.R., *Mol. Microbiol.*, 1996, vol. 19, no. 1, pp. 125–137.
4. McGuire, A.M., De Wulf, P., Church, G.M., and Lin, E.C.C., *Molecular Microbiology*, 1999, vol. 32, no. 1, pp. 219–221.
5. Favorov, A.V. and Gerasimova, A.V., Proceedings of the Moscow Conference on Computational Molecular Biology MCCMB'03, 2003, pp. 67–69.
6. Mironov, A.A., Vinokurova, N.P., and Gelfand, M.S., *Molecular Biology*, vol. 34, no. 2, pp. 222–231.
7. Gerasimova, A.V., Rodionov, D.A., Mironov, A.A., and Gelfand, M.S., *Molecular Biology*, vol. 35, no. 6, pp. 853–861.
8. Blattner, F.R., Plunkett, G., Bloch, C.A., *et al*., *Science*, 1997, vol. 77, pp. 1453–1474.
9. Parkhill, J., Wren, B.W., Thomson, N.R., *et al*., *Nature*, 2001, vol. 413, no. 6855, pp. 523–527.
10. May, B.J., Zhang, Q., Li, L.L., *et al*., *Proc. Natl. Acad. Sci. USA*, 2001, vol. 98, no. 6, pp. 3460–3465.
11. Kim, Y.R., Lee, S.E., Kim, C.M., *et al*., *Infect. Immun*., 2003, vol. 71, no. 10, pp. 5461–5471.
12. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., *et al*., *Nucl. Acids Res*., 2000, vol. 28, pp. 15–18.
13. Mironov, A.A. and Gelfand, M.S., *Molecular Biology*, vol. 33, no. 1, p. 109.

14. Compan, I. and Touati, D., *Mol. Microbiol*., 1994, vol. 11, no. 5, pp. 955–964.

15. Lynch, A.S. and Lin, E.C., *J. Bacteriol.*, 1996, vol. 178, no. 21, pp. 6238–6249.

16. Pellicer, M.T., Fernandez, C., Badia, J., *et al*., *J. Biol. Chem*., 1999, vol. 274, no. 3, pp. 1745–1752.

17. Hassan, H.M. and Sun, H.C., *Proc. Natl. Acad. Sci. USA*., 1992, vol. 89, no. 8, pp. 3217–3221.

18. Pellicer, M.T., Lynch, A.S., De Wulf, P., *et al*., *Mol. Gen. Genet*., 1999, vol. 261, no. 1, pp. 170–176.

19. Cunningham, L. and Guest, J.R., *Microbiology*, 1998, vol. 144, no. 8, pp. 2113–2123.

20. Chao, G., Shen, J., Tseng, C.P., *et al*., *J. Bacteriol*., 1997, vol. 179, no. 13, pp. 4299–4304.

21. Tseng, C.P., Albrecht, J., and Gunsalus, R.P., *J. Bacteriol*., 1996, vol. 178, no. 4, pp. 1094–1098.

22. Campbell, J.W., Morgan-Kiss, R.M., and Cronan, J.E.Jr., *Mol. Microbiol*., 2003, vol. 47, no. 3, pp. 793–805.

23. Kaiser, M. and Sawers, G., *Microbiology*, 1997, vol. 143, Pt. 3, pp. 775–783.

24. Wyborn, N.R., Messenger, S.L., Henderson, R.A., *et al*., *Microbiology*, 2002, vol. 148, no. 4, pp. 1015–1026.

25. Quail, M.A., Haydon, D.J., and Guest, J.R., *Mol. Microbiol*., 1994, vol. 12, no. 1, pp. 95–104.

26. van der Rest, M.E., Frank, C., and Molenaar, D., *J. Bacteriol*., 2000, vol. 182, no. 24, pp. 6892–6899.

27. Colloms, S.D., Alen, C., and Sherratt, D.J., *Mol. Microbiol*., 1998, vol. 28, no. 3, pp. 521–530.

28. Wood, J.M., *Proc. Natl. Acad. Sci. USA*, 1987, vol. 84, no. 2, pp. 373–377.

29. Ritz, D., Patel, H., Doan, B., *et al*., *J. Biol. Chem*., 2000, vol. 275, no. 4, pp. 2505–2512.

30. Chen, P., Andersson, D.I., and Roth, J.R., *J. Bacteriol*., 1994, vol. 176, no. 17, pp. 5474–5482.

31. Matsubara, M., Kitaoka, S.I., Takeda, S.I., *et al*., *Genes Cells*, 2000, vol. 5, no. 7, pp. 555–569.

32. Golby, P., Kelly, D.J., Guest, J.R., *et al*., *J. Bacteriol*., 1998, vol. 180, no. 24, pp. 6586–6596.

33. Lynch, T.W., Read, E.K., Mattis, A.N., Gardner, J.F., and Rice, P.A., *J. Mol. Biol*., 2003, vol. 330, no. 3, pp. 493–502.