

UDC 577.218.577.577.121.9

# Study of Regulation of Long-Chain Fatty Acid Metabolism Using Computer Analysis of Complete Bacterial Genomes

N. S. Sadovskaya<sup>1</sup>, O. N. Laikova<sup>2</sup>, A. A. Mironov<sup>2</sup>, and M. S. Gelfand<sup>2</sup>

<sup>1</sup> Institute for Information Transmission Problems, Moscow, 101447 Russia

<sup>2</sup> State Research Center for Biotechnology GosNIIGenetika, Moscow, 113545 Russia;  
E-mail: misha@imb.imb.ac.ru

Received April 18, 2001

**Abstract**—One of the main trends in the prokaryote genomics is the comparative analysis of metabolic pathways. This method can be used for the analysis of experimentally studied systems of co-regulated genes, as well as genes with unknown regulatory signals. In this study we apply the comparative analysis of regulatory signals to the genes of the enzymes for fatty acid metabolism from *Escherichia coli*, *Haemophilus influenzae*, *Vibrio cholerae*, and *Yersinia pestis*. Transcription of these genes is regulated by the FadR protein. We describe the FadR regulation of long-chain fatty acid oxidation and partial regulation of fatty acid biosynthesis. We also demonstrate that the gene *yafH* encoding acyl-CoA dehydrogenase is identical to the gene *fadE* previously identified by genetic techniques.

**Key words:** computer analysis, functional signals, FadR, fatty acids, *Escherichia coli*, *Haemophilus influenzae*, *Vibrio cholerae*, *Yersinia pestis*

## INTRODUCTION

The wild-type *Escherichia coli* can use fatty acids (FA) with the chain length of 12 and more carbon atoms (long-chain FA) as the sole source of carbon. The cells start growing after a lag period that is necessary for induction of the fad regulon (genes regulated by FadR). Both the wild-type *E. coli* and strains constitutively expressing the fad regulon as a result of mutations in the *fadR* gene [1] can use fatty acids with the chain length of 7–11 carbon atoms (medium-chain FA) only after induction of the fad regulon by long-chain FA. Catabolism of FA with the chain length of 4–6 requires not only enzymes of the fad system, but also two degradation enzymes encoded by genes *atoD*, *atoA*, and *atoB*. These genes are positively regulated by the *atoC* gene product. Biosynthesis of FA is regulated by FadR and the repressor FabR [2]. Here we consider only the FadR-regulated genes.

The FadR protein represses the long-chain FA oxidation and activates several stages of the FA biosynthesis. It is known that FadR binding increases 20-fold the *fabA* promoter activity [3]. Genes of the fad regulon enzymes are randomly dispersed in the *E. coli* chromosome, except for *fadB* and *fadA* which form an operon. The fad regulon is responsible for transport (*fadL*) of long-chain FA and also for activation (*fadD*) and oxidation (*fadE*, *fadH*, and *fadBA* [4]) of long-chain and medium-chain FA. In addition, FadR acti-

vates two genes for the FA biosynthesis (*fabA* and *fabB*) and the gene *iclR*, which regulates the enzymes of the glyoxylate pathway [5].

The aim of this study was to characterize the fad regulon in *E. coli*, *Haemophilus influenzae*, *Vibrio cholerae*, *Yersinia pestis* and to find unknown members of the fad regulon.

## METHODS

The comparative approach to the analysis of regulation is based on the assumption that relative genomes have identical structure of the regulon. Thus, true regulatory sites occur upstream of the orthologous genes, and false sites (overprediction) are randomly dispersed. Therefore, a pair of genes, one from each genome, can be included into a regulon if:

(i) these genes are orthologs, i.e., they are homologous, and their divergence is due to speciation rather than duplication (thus their cell function is most probably conserved);

(ii) there are candidate sites upstream of these genes in the genomes under consideration [6].

As mentioned above, FadR activates genes *fabA*, *fabB* [7], and *iclR* [5] and represses *fadL*, *fadD*, and *fadB* [1, 4]. Thus, we have selected these genes in order to construct the training sample. Positional

weights of nucleotides were derived from the standard formula [8]

$$W(b, k)$$

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

where  $N(b, k)$  is the count of nucleotide  $b$  at position  $k$ . The score of a candidate site is the sum of position weights of the constituent nucleotides. The base of the logarithm is chosen such that the scores of random oligonucleotides follow the standard Gaussian distribution with zero mean and unit variance. The nucleotide weight matrix for candidate FadR-binding sites is:

<i>a</i>	<i>c</i>	<i>g</i>	<i>t</i>
0.24	-0.15	-0.04	-0.04
0.28	-0.35	0.07	0.00
-0.11	0.33	-0.11	-0.11
0.19	-0.28	-0.28	0.37
0.00	0.07	0.28	-0.35
0.26	-0.29	0.32	-0.29
-0.24	-0.24	0.00	0.47
0.09	0.43	-0.26	-0.26
-0.08	0.08	0.08	-0.08
-0.26	-0.26	0.43	0.09
0.47	0.00	-0.24	-0.24
-0.29	0.32	-0.29	0.26
-0.35	0.28	0.07	0.00
0.37	-0.28	-0.28	0.19
-0.11	-0.11	0.33	-0.11
0.00	0.07	-0.35	0.28
-0.04	-0.04	-0.15	0.24.

The recognition rule thus derived was used to scan the potential regulatory regions of the *E. coli*, *H. influenzae*, *V. cholerae*, and *Y. pestis* genes. We considered regions (-200 to +50) in *E. coli* and (-250 to +100) in *H. influenzae*, *V. cholerae*, and *Y. pestis*. These positions were set relative to the translation start of each gene. The palindromic candidate FadR-binding sites, their scores, and positions relative to the first gene in the operon are shown in the table. A site was included in the table only if there were potential sites in orthologous operons (see (i) and (ii) above). We also considered strong (weight >4.00) and weak (4.00 > weight > 3.90) candidate regulatory sites in the genes functionally related to the metabolism of fatty acids. Initially, we found 27 strong candidate sites in the *E. coli* genome, 4 in the *H. influenzae* genome, 23 in *V. cholerae*, and 22 in *Y. pestis*. We also found 9 weak candidate sites in the *E. coli* genome, 7 in *H. influenzae*, 10 in *V. cholerae*, and 8 in *Y. pestis*. It is obvious that this number of candidate sites cannot be acceptable.

Therefore, we have applied the additional criteria described above.

All calculations were done using the software package Genome Explorer [9].

Genes were functionally annotated using BLASTA (<http://www.ncbi.nlm.nih.gov/BLASTA/>) [10] and the data bank of amino acid sequences SWISS-PROT (<http://www.expasy.hcuge.ch/sprot/>) [11]. Search for orthologous genes was done using the database COG (<http://www.ncbi.nlm.nih.gov/COG/>) [12]. Intergene regions in relative organisms were aligned using the program ClustalX [13]. Metabolic pathways were analyzed using the database KEGG (<http://www.genome.ad.jp/kegg/kegg2.html/>) [14].

## RESULTS AND DISCUSSION

The scheme of long-chain FA oxidation and the genes of the catalyzing enzymes are shown in the figure.

Note that the enzyme FadR does not regulate expression of its own gene in *E. coli*, *H. influenzae*, *V. cholerae*, and *Y. pestis*.

Ten operons regulated by FadR were found in the *E. coli* genome. Consider them in more detail.

It is known that the genes of the enzymes for utilization (*fadL*, *fadD*, *fadBA*) and synthesis (*fabA*, *fabB*) of long-chain FA [7], and the repressor of the glyoxylate pathway (*iclR*) [5] are all controlled by *fadR*. Therefore, they should have a strong site in the upstream regions. This assumption was confirmed by the results obtained: the score of a site ranged from 4.41 to 5.19 for these genes.

Analysis of the operon *b2342-41* corroborated that it is a paralog of *fadBA*. We identified a strong FadR-binding site upstream of this operon. Apparently, the function of *b2342-41* is similar to that of *fadBA*.

The *yafH* gene encoding acyl-CoA dehydrogenase also has a strong (4.09) site in the regulatory region. In addition, there are orthologs of this gene retaining strong candidate sites in the genomes of *V. cholerae* and *Y. pestis*. It is known that the *fadE* gene encodes acyl-CoA dehydrogenase [1] of the long-chain FA  $\beta$ -oxidation pathway. Thus, *yafH* is *fadE*. Note that formal analysis of protein similarity, for example, using the COG system, does not allow one to select *fadE* from a family of related genes, namely *ydiO*, *caiA*, *aidB*, and *yafH*.

The *fadH* gene encodes 2,4-dienoyl-CoA reductase, an enzyme catalyzing one stage of oxidation of unsaturated long-chain FA [15, 16]. Furthermore, there is a candidate FadR-binding site upstream of this gene. Thus we conclude that *fadH* is regulated by FadR. This is in good agreement with the fact that

Known and predicted FadR signals in the genomes of *Escherichia coli*, *Haemophilus influenzae*, *Vibrio cholerae*, and *Yersinia pestis*

Gene	Position	Weight	Signal	Gene	Position	Weight	Signal
<i>Escherichia coli</i>				<i>Haemophilus influenzae</i>			
<i>fadL</i>	-136	4.41	AgCTGGTCCGACcTaTa	<i>HI0401</i>	-135	3.39	AACTaGTCGtAgCtcTa
<i>fadD</i>	-170	4.54	AgCTGGTatGAtgAGTT	<i>HI0390.1</i>	-165	2.65	cttTGGTatGttCAGcc
<i>fadBA</i>	-109	4.62	AtCTGGTaCGACCAGaT	#	#	#	#
<i>pdhR</i>	-48	4.14	AAAtTGGTaaGACCAaTT	#	#	#	#
<i>b0221 (fadE)</i>	-38	4.09	AAgTGGTCaGACCtccT	#	#	#	#
<i>fabA</i>	-72	5.00	AACTGaTCGGACttGTT	<i>HI1325</i>	-121	4.00	gACTGcTCCGACaAGTT
<i>fabB</i>	-82	4.51	ggCTGaTCGGACttGTT	<i>HI1533</i>	-95	3.37	AACTGGTTCGaACaAaTg
<i>b2342-41</i>	-44	4.46	AtCaGGTCaGACCacTT	#	#	#	#
<i>fadR</i>	-51	3.80	ctCTGGTatGAtgAGTc	<i>HI0426</i>	-23	2.43	tttTtaTCtGAtttTa
<i>ygjL (fadH)</i>	-46	4.53	AACTcaTCCGACCAcaT	#	#	#	#
<i>iclR</i>	-86	5.19	AACTcaTCGGAtCAGTT	#	#	#	#
<i>Vibrio cholerae</i>				<i>Yersinia pestis</i>			
<i>VC1043</i>	-135	3.45	cAaTGGTCCGAttcTa	<i>fadL</i>	-133	4.05	cACaGGTCCGACcTaTa
<i>VC1985</i>	-180	3.18	AgCaccTCGGctgtGcT	<i>fadD</i>	-171	4.28	AACTGGTaaGctgAGTT
<i>VC2758-59</i>	-119	4.11	AACTGGTCaaACCAGaa	<i>fadBA</i>	-110	4.46	AtCTGGTCatACCAGaT
<i>VC2415</i>	-54	3.80	AAAtTGGTattACCAaTT	<i>pdhR</i>	-68	3.80	AAAtTGGTattACCAaTT
<i>VC2231</i>	-83	3.95	AACTGGTtaGACCacTa	<i>b0221 (fadE)</i>	-36	4.35	AAcAGGTCaGACCtccT
<i>VC1483</i>	-70	4.00	cACTGaTCGGAgttGTT	<i>fabA</i>	-74	4.02	ggCTaaTCGGACttGcT
<i>VC2019</i>	-80	3.00	gggaaagCtGACCacTT	<i>fabB</i>	-75	4.40	cgCTGaTCGGACttGTT
<i>VC1046-47</i>	-106	3.29	AAaaaaTCctACCAaca	<i>b2342-41</i>	-66	4.72	AtCaGGTCaGACCtGTT
<i>VC1900</i>	-58	3.64	tAgTGGTatGAtgAGTg	<i>fadR</i>	-59	3.59	gtCTGGTatGAtgAGcg
<i>VC1993</i>	-42	4.08	ttCTGGTCaGACCAtaT	<i>ygjL (fadH)</i>	-84	4.53	AtCTcaTCCGACCAcTT
#	#	#	#	<i>iclR</i>	-109	2.67	AAcAGGcgGGAttAccc

Note: Candidate FadR sites, their positions, and their scores are given for each gene of the *fad* regulon. The sites with a score less than 4.00 are weak, and the sites with a score less than 3.90 are not significant; # denotes the absence of the gene. The names of open reading frames of *Yersinia pestis* are provisional.

FadR functions as a repressor of the long-chain FA degradation.

The results obtained imply that FadR regulates all stages of the long-chain FA oxidation and partially regulates the FA biosynthesis.

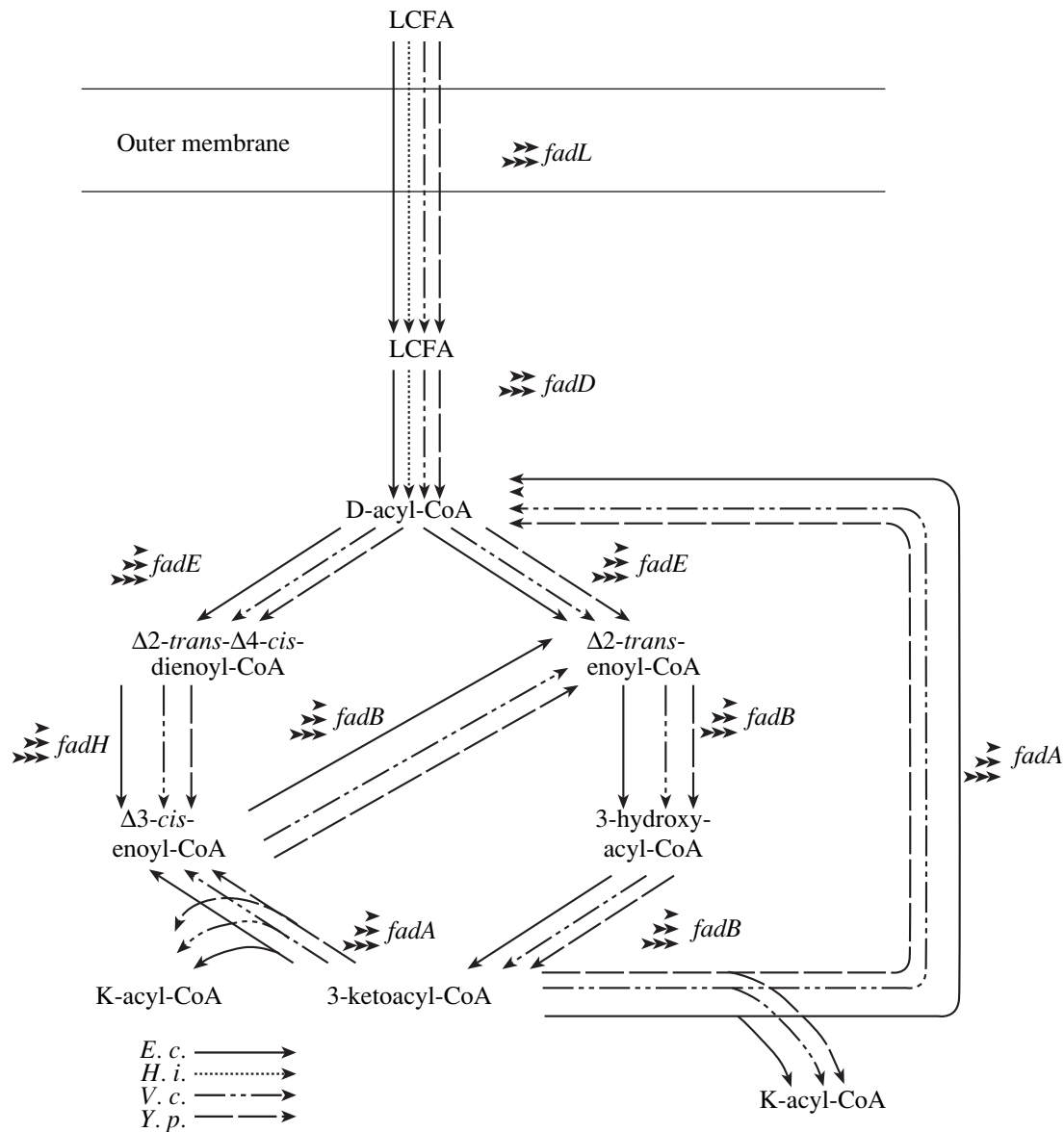
In addition, a candidate FadR-binding site was observed upstream of gene *pdhR*, which encodes the transcription regulator of the pyruvate dehydrogenase complex. PdhR has its own strong regulatory signal [17]. The *pdhR* autoregulation site appeared to be also a strong FadR site, but this is most probably just a coincidence. Note that there are orthologs of *pdhR* in the genomes of *V. cholerae* and *Y. pestis*, and *V. cholerae* has a strong FadR-binding site upstream of the ortholog.

The systems of *E. coli* and *Y. pestis* are almost the same. All operons are conserved, and sites are lost only upstream of *fadR* and *iclR*. This implies that the

metabolic pathways of long-chain FA and their regulation are identical in *Y. pestis* and *E. coli*.

Comparison between *E. coli* and *V. cholerae* demonstrated that the *V. cholerae* genome includes orthologs of all genes except *iclR*. However, only three genes have strong upstream candidate FadR-binding sites, namely *fabA*, *fadB*, and *fadH*, and there is a strongly degenerate site upstream of *ygjL (fadE)*. In addition, the *fabB* gene is likely to be a pseudogene [18], because its reading frame is shifted. The remaining genes can be constitutive or controlled by other factors.

The genome of *H. influenzae* lacks most of the genes for the long-chain-FA oxidation enzymes. We have found only orthologs of *fadL*, *fadD*, *fabA*, *fabB*, and *fadR*, and a regulatory site upstream of *fabA*. Apparently, the expression of the remaining genes is constitutive or controlled by another gene. However,



Oxidation of long-chain fatty acids of gamma-proteobacteria *E. coli*, *H. influenzae*, *V. cholerae*, *Y. pestis*. The presence of candidate FadR-binding sites in the regulatory region of *E. coli* (>>>), *Y. pestis* (>>), *V. cholerae* (>). Boldface indicates strong sites, and the standard type indicates weak sites. We have not found any candidate sites regulating genes of the FA oxidation system in the genome of *H. influenzae*.

we have not discovered any conserved signals in their regulatory regions. The absence of a large number of genes implies that there is almost no long-chain FA metabolism in *H. influenzae*.

Thus, the comparison between *fad* regulons in several gamma-proteobacteria revealed their low conservation. The *fad* regulons of *E. coli* and *Y. pestis* are almost identical. *V. cholerae* retains most genes, but they are not FadR-regulated. Most genes of the *fad* regulon are absent from *H. influenzae*. Using the comparative analysis, we managed to find new members of this regulon in the *E. coli* genome (*b2342-42*), to iden-

tify the *fadE* gene, and to predict the FadR-binding site upstream of the *fadH* gene.

#### 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. Clark, D.P. and Cronan, J.E. Jr., *Escherichia coli* and *Salmonella*. *Cellular and Molecular Biology*,

- Neindhard, F.C., Ed., Washington DC: ASM Press, 1996, pp. 343–357.
2. McCue, L.A., Thompson, W., Carmark, C.S., *et al.*, *Nucleic Acids Res.*, 2001, vol. 29, pp. 774–782.
  3. Cronan, J.E., Jr. and Subrahmanyam, S., *Mol. Microbiol.*, 1998, vol. 29, pp. 937–943.
  4. Raman, N., Black, P.N., and DiRusso, C.C., *J. Biol. Chem.*, 1997, vol. 272, pp. 30645–30650.
  5. Gui, L., Sunnarborg, A., and Laporte, D.C., *J. Bacteriol.*, 1996, vol. 178, pp. 4704–4709.
  6. Gelfand, M.S. and Mironov, A.A., *Mol. Biol.*, 1999, vol. 33, pp. 772–778.
  7. Cronan, J.E., Jr. and Rock, C.O., *Escherichia coli and Salmonella. Cellular and Molecular Biology*, Neindhard, F.C., Ed., Washington, DC: ASM Press, 1996, pp. 612–636.
  8. Mironov, A.A. and Gelfand, M.S., *Mol. Biol.*, 1999, vol. 33, pp. 127–132.
  9. Mironov, A.A., Vinokurova, N.P., and Gelfand, M.S., *Mol. Biol.*, 2000, vol. 34, pp. 253–262.
  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. Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., *et al.*, *Nucleic Acids Res.*, 2001, vol. 29, pp. 22–28.
  13. Heringa, J., *Comput. Chem.*, 1999, vol. 23, pp. 341–364.
  14. Wixon, J. and Kell, D., *Yeast.*, 2000, vol. 17, pp. 48–55.
  15. He, X.Y., Yang, S.Y., and Schulz, H., *Eur. J. Biochem.*, 1997, vol. 248, pp. 516–520.
  16. You, S.Y., Cosloy, S., and Schulz, H., *J. Biol. Chem.*, 1989, vol. 264, pp. 16489–16495.
  17. Quail, M.A. and Guest, J.R., *Mol. Microbiol.*, 1995, vol. 15, pp. 519–529.
  18. Heidelberg, J.F., Eisen, J.A., Nelson, W.C., *et al.*, *Nature*, 2000, vol. 406, pp. 469–470.