Journal of Bioinformatics and Computational Biology Vol. 3, No. 4 (2005) 1007–1019 © Imperial College Press



# EVOLUTION OF THE NADR REGULON IN ENTEROBACTERIACEAE

ANNA V. GERASIMOVA\*

Laboratory of Bioinformatics, State Scientific Center GOSNIIGenetika 1-iy Dorozhny proezd 1, Moscow, 113545, Russia a\_gerasimova@yahoo.com

### MIKHAIL S. GELFAND

Institute for Information Transmission Problems, RAS Bolshoy Karetny per. 19, Moscow, GSP-4, 127994, Russia gelfand@iitp.ru

> Received 5 February 2005 Revised 18 February 2005 Accepted 24 February 2005

The NAD biosynthetic pathway and NAD transformations in *E. coli* and *S. typhi* are well characterized. Using comparative genomics methods we describe the NadR regulon in other *Enterobacteriaceae*, identity new candidate regulon members and demonstrate that even a very simple regulon covering an essential methabolic pathway could be different in closely related genomes.

*Keywords*: NAD biosynthesis; NadR; transcription factor; regulation of transcription; comparative genomics; phylogenetic footprinting; evolution.

## 1. Introduction

The comparative approach to the analysis of regulation is based on the assumption that regulons are conserved in related bacteria containing ortologous transcription factors.

This approach, reviewed in Refs. 1–3, has been successfully applied to the analysis of many regulatory systems<sup>4–15</sup> and served as a base for large-scale analyses of regulation in all prokaryotes,<sup>16,17</sup> as well as selected taxonomic groups of gamma-proteobacteria,<sup>18,19</sup> delta-proteobacteria,<sup>20</sup> and gram-positive bacteria,<sup>21,22</sup> resulting in identification of numerous new signals and functional annotation of tens of hypothetical genes. Many of such predictions were subsequently confirmed in experiment,<sup>23,24,12</sup> or even served as a starting point for experimental

 $^{*}$ Corresponding author.

analysis.<sup>18,25–27</sup> There exist several Internet servers for comparative analysis of bacterial regulation, in particular, EnteriX<sup>28</sup> and PredictRegulon.<sup>29</sup>

In an attempt to analyze the evolutionary dynamics of a relatively simple, wellstudied regulon that includes genes from an essential part of the metabolism, we considered the NadR regulon in *Enterobacteriaceae*.

The nicotinamide adenine dinucleotides (NAD, NADH, NADP, NADPH) are essential cofactors in all living systems and function as hydride acceptors (NAD, NADP) and donors (NADH, NADPH) in biochemical redox reactions.<sup>30</sup> At high internal levels of NAD, the transcriptional regulator NadR represses the *de novo* synthesis and salvage pathways. NadR is a multifunctional protein, consisting of an N-terminal DNA-binding domain which represses NAD biosynthesis, a central nicotinamide mononucleotide adehyltransferase (NMNAT) domain and a Cterminal RNK domain.<sup>31,32</sup>

The NAD biosynthetic pathway and transformations are shown in Fig.  $1.^{31}$ 

Genes known to be repressed by NadR in *E. coli* and *S. typhi* are marked by rectangles. These are two NAD biosynthesis genes, nadA and nadB, and a niacin salvage gene pncB.<sup>32,33</sup>

### 2. Data and Methods

The complete genomes of Escherichia coli K-12 MG1655<sup>34</sup> (EC), Shigella flexneri 2457T<sup>35</sup> (SF), Salmonella typhi CT18<sup>36</sup> (ST), Erwinia carotovora subsp. atroseptica SCRI1043<sup>37</sup> (ERW), Yersinia pestis CO92<sup>38</sup> (YP) and Photorhabdus luminescens subsp. laumondii TT01<sup>39</sup> (PHL) were obtained from Genbank.<sup>40</sup>



Fig. 1. The NAD biosynthetic pathway and transformations in *Enterobacteriaceae*. Notation: "\*": enzymatic domain; "#": NMN transporter, regulated within the *nadApnuC* operon.

Incomplete genomes of *Klebsiella pneumoniae* MGH78578 (KP) and *Serratia marcescens* Db11 (SM) were downloaded from the websites of the Washington University Consortium (www.genome.wustl.edu), and *Yersinia enterocolitica* 8081 (YE), from the Sanger Institute website (www.sanger.ac.uk).

Profiles (positional weight matrices) for the identification of candidate NadRbinding sites were constructed using SignalX.<sup>4</sup> The training set consists of upstream regions of *nadA* from *E. coli*, *S. typhi* and *Y. pestis*, *nadB* from *E. coli* and *S. typhi*, and *pncB* from *E.coli*, *S.typhi* and *Y. pestis*.

Sequence logo was constructed using WebLogo.<sup>41</sup> Orthologs were identified by the bidirectional best hits criterion<sup>42</sup> and, if necessary, verified by construction of phylogenetic trees using PHYLIP.<sup>43</sup> Multiple nucleotide and protein alignments were constructed using ClustalX.<sup>44</sup> Genome analyses were performed using GenomeExplore.<sup>45</sup>

## 3. Results and Discussion

NadR orthologs were identified in all studied Enterobacteria. Multiple protein alignment demonstrated that NadR orthologs in all considered genomes contained DNAbinding domain, NMNAT domain and RNK domain.

It is known that in some gamma-proteobacteria, for example in Haemophilusinfluenzae, NadR orthologs do not contain the DNA-binding domain<sup>31</sup> and thus have only enzymatic, but not regulatory role. Indeed, no DNA-binding domains were found in NadR orthologs from genomes outside the *Enterobacteriaceae* and *Pasteurellaceae* families. Among the latter, *Haemophilus influenzae* is the only genome with NadR lacking the DNA-binding domain. NadR of other *Pasteurellaceae* have the DNA-binding domain, but these genomes have no *nadA*, *nadB* and *pncB* orthologs, nor do they have candidate sites for the enterobacterial NadR-signal. Thus here we restricted the analysis to the NadR regulon in *Enterobacteriaceae*.

The recognition profile was constructed as described above. The sequence logo of the NadR signal is shown in Fig. 2.



Fig. 2. Sequence logo of NadR-sites from the training set. The total height of the symbols in each position equals the positional information content, whereas the height of individual symbols is proportional to the positional nucleotide frequency, with the most frequent nucleotide shown at the top.

	nadB		nadA		Total		10001		ynill/m		rpsr	
Genome	Name of Ortologues Gene	Score	Name of Ortologues Gene	Score	Name of Ortologues Gene	Score	Name of Ortologues Gene	Score	Name of Ortologues Gene	Score	Name of Ortologues Gene	Score
EC	nadB	6.21	nadA	5.95	pncB	5.63	nadR		ynfL/M	4.69	rpsP	
$_{\rm SF}$	in DNA	6.21	nadA	5.95	pncB	5.63	nadR		${ m ynfL/M}$	4.69	rpsP	
$\mathbf{TS}$	STY2834	6.21 5.09	m 27V072	5.95	STY1010	6.06	nadR		$\mathrm{STY1578}/79$		STY2863	
KP	$\operatorname{nadB}$	6.21	nadA	5.95	pncB	5.11	nadR		${ m ynfL/M}$	4.69	in DNA	
ERW	nadB		ECA1378	4.62 5.23	$\operatorname{pncB}$		ECA0463	5.62	ECA2259/60	4.69	ECA3359	5.10
SM	nadB		nadA	5.17 5.81	$\operatorname{pncB}$	5.07	nadR	5.71	${ m ynfL/M}$	4.69	in DNA	5.16
ΥP	nadB		nadA	4.29 5.86	$\operatorname{pncB}$	4.62	nadR	5.91	in DNA/YPO2266	5.33	rpsP	5.16
YE	RYE01420		RYE03344	5.86	RYE02025		RYE00967	5.63	$\mathrm{RYE00573}/74$	4.69	RYE01243	5.16
PHL	$\operatorname{nadB}$		plu1468	6.43	pcnB		nadR		plu2225/24	5.33	rpsP	4.80

Table 1. Genes from candidate NadR regulons.

1010 A. V. Gerasimova & M. S. Gelfand

The signal is a palindrome with six conserved positions at each side and a spacer of six relatively less conserved positions.

The study started with identification of orthologs of genes that constitute the NadR regulation in  $E. \ coli$  and analysis of their regulation. The results are shown in Table 1.

NadR-sites of the nadA genes are conserved and they form the only conserved island in the alignment of upstream regulons (Fig. 3).

Additional candidate sites were identified in S. marcescens E. carotovora.

EC  SF  ST  Plu  YP  YE  ERW  SM	TGTAAACCTGATGACATCGTCAGAGCTTACTGTGCAAGCAA
EC  SF  KP  ST  Plu  YP  YE  ERW  SM	TAGGCGTAAAATGACGCATCCTGCACATTAGGCGTAATTCGA-GTGACTTTTCCCCA TAGGCGTAATTCCGA-GTGACTTTTCCCCA GCGCAGCACCGTCCCCGGGAAAAGGGTCTGATGTACAAAATTCATCACGACTTTTAGCTC AAGACGTTACGCCGCCATCAGATACGGACTACGTAATT-ATTGTGACTTTCCTTAT TCTGTTTAGCCATCCTATCATAGATAAAATATCTCGCGACTATTAACTA TTAACTCTATTAAACATTTTCGCTATTAATCGACAAAATTTCGGCGACTTTTTGCCA TTAACTCTATTAAACATTTTTGCCATTTTAGACAATATCTGGCGACTTTTCGCCA <b>TAACGATAGTAAACA</b> AAGTGGTGGGTTTTTTGGAAAAAACACCGCGACTTTTAGGCG <b>TTAATCTATTAAACA</b>
	тдттта таласа
EC  SF  ST  Plu  YP  YE  ERW  SM	CCATTCGACTATCT <b>TGTTTA</b> GCATATA <b>AAACA</b> AATTACACCGA-TAACAGCGAATAT CCATTCGACTATCT <b>TGTTTA</b> GCATATA <b>AAACA</b> AATTACACCGA-TAACAGCGAATGT GTAATCGGCTATCT <b>TGTTTA</b> GCATATA <b>AAACA</b> CCATGACCGTA-TTAAGCCGCCGAC TGAATCAGCTATCT <b>TGTTTA</b> GCATATA <b>AAACA</b> AATTGACCG-A-TTGTGGCGTTTAT TAAATTGATCATTT <b>TGTTTA</b> GGATAAA <b>AAACA</b> AATCTGGGCAAATTTT-CGGACTGA AATATTGGTTATTT <b>TGTTTA</b> GGATAAA <b>AAACA</b> AATCAAATCAAATCCTAC-CTGGCAGCCGAC TGTTTTCGGTATTT <b>TGTTTA</b> GGATAAA <b>AAACA</b> AATCAAATCAAATCCTGACTGGCAGCCTGA TGTTTTCGGTATTT <b>TGTTTA</b> GGATAAA <b>AAAC</b> AAATCAAATCAAATCCAATCCTGACTGGCAGCCTGA TGTTTTCGGTATTT <b>TGTTTA</b> GGATAAA <b>AAAC</b> AAATCAATCCAATCCTGACTGGCAGCCTGA TGTTTTCGGTATTT <b>TGTTTA</b> GGATAAA <b>AAACT</b> CCCGATGCTGATAG
EC  SF  ST  Plu  YP  YE  ERW  SM	TACGC - TAATGTCGGTTTTA ACGTTAAGCCTGTAAAACGAGA TGGTAAGATGAGCG       TACGC - TAATGTCGGTTTTA ACGTTAAGCCTGTAAAACGAGA TGGTAAGATGAGCG       AGGCCATACTGACGGTTATA GAGTTAAGCCAGTAAAACGAGA TTGCATGATGAGCG       TACGC - TAATGTCGGTTGT TCGTTAAGCCAGTAAAACGAGA TTGCATGATGAGCG       AAACCAAAATTATCGGTTGT TCGTTAAGTCAGTAAAACGAGAGTCAAAGTAATGAGCG       GCAGCGATCTGATAGCTTCAGTAATACCAAGTCAAACGAGAGTCAAAGTAATGAGCG       GCAGCGATCTGATAGCTTCAGTAATACCACAGC AACAGAGA - TGTGAGCGATGAGCG       AAGGCA TTATGGCTCCG GAATCACTGGC AATAGAGA - TGTGAGCGATGAGCG       TAGGCCATATCCAGGTGTTA - CAATCCGAGTTGTCATCGTGGAA - ATCCA - TAATGAATA       TAGGCGATATTCACGCGGCA - GGCGCCCCAGC - CTAGCATGAGAA - TTGCAGCGATGAGTG       *     *     *     *     *     *     *     *     *

Fig. 3. Conservation of NadR-sites upstream of *nadA*. The sites are shadowed; positions conforming to the signal consensus and start codons (ATG) are set in boldface.

#### 1012 A. V. Gerasimova & M. S. Gelfand

EC	TAACCCAACG	GCCTTTTTA	ATTTCACC	ACCTAATC	CTCCACCA	AGC	-CAGTAACT
SF	TAACCCAACG	GCCTTTTTA	ATTTCACC	ACCTAATC	CTCCACCA	AGC	-CAGTAACT
ST	TAACCTAACA	GCATCTTT	TTTCACT	ACAAAATC	CGACGCTA	ACACCCTGC	CCTATAAAA
КР	TATCGTAACA	CGCCGTTTA	TTTCACT	ATAAAATC	CAATGCC	ATCAACCTTC	CCCGCGTCT
1	** * ***	* * * *	*****	* ****	* * *	* *	*
EC	<u> </u>			CCCTGCGT	САСССТСТ	пппассаася	GTAACAAAT
CEL		۳					
Curl Curl							
SI	TATTTTTGC	CGITIAICI			CTTTAIG		ACAACACTT
КР	CA'I"I"I"I'CAGC	GCGCAAGAG	CGCCGIIIII	CCGTTCGC	C'I''I''I'-'I'G'I	I"I"PAGCCGPC	ACAACAGCA
	* **		* *	*	* * *	****	* * * *
							TGTTTA
EC	ATTAAAATAG	CAGGTGTT	TATCCGCA	CAACATGA	TGCTATGO	CTGACCAAAC	<b>TGTTTA</b> CCA <b>TGTTTA</b> G
EC   SF	ATTAAAATAG ATTAAAATAG	CAGGTGTT CAGGTGTT	PATCCGCA PATTCGCA	.CAACATGA .CAACATGA	TGCTATG( TGCTATG(	CTGACCAAAC CTGACCAAAC	<b>TGTTTA</b> CCA <b>TGTTTA</b> G CAA <b>TGTTTA</b> G
EC   SF   ST	АТТААААТАG АТТААААТАG ТТGАААТСАТ	CAGGTGTT CAGGTGTT AACGTGCTT	PATCCGCA PATTCGCA PTTTAGCG	.CAACATGA .CAACATGA :CCATATAG	TGCTATG( TGCTATG( TGCTAAT(	CTGACCAAAC CTGACCAAAC CTGCCGCAAC	<b>TGTTTA</b> CCA <b>TGTTTA</b> G CAA <b>TGTTTA</b> G CCA <b>TGTTTA</b> G
EC   SF   ST   KP	АТТААААТАG АТТААААТАG ТТGАААТСАТ GACAAAA-AA	CAGGTGTT CAGGTGTT AACGTGCT ATTGTACGZ	PATCCGCA PATTCGCA PTTTAGCG	CAACATGA CAACATGA CCATATAG	TGCTATGO TGCTATGO TGCTAATO TGCTAATTO	CTGACCAAAC CTGACCAAAC CTGCCGCAAC	TGTTTA CATGTTTAG CATGTTTAG CATGTTTAG
EC   SF   ST   KP	ATTAAAATAG ATTAAAATAG TTGAAATCAT GACAAAA-AA *** *	CAGGTGTT CAGGTGTT AACGTGCT ATTGTACG4 * *	PATCCGCA PATTCGCA PTTTAGCG ATTCCTCA * *	CAACATGA CAACATGA CCATATAG CGGACCGG *	TGCTATG( TGCTATG( TGCTAAT( TGCTATT( * * * * *	CTGACCAAAC CTGACCAAAC CTGCCGCAAC GTGAGCTAAA	TGTTTA CATGTTTAG CATGTTTAG CATGTTTAG ATG <mark>TGTTTA</mark> G
EC   SF   ST   KP	АТТААААТАG АТТААААТАG ТТGАААТСАТ GACAAAA-AA *** *	CAGGTGTT CAGGTGTT AACGTGCT ATTGTACG **	TATCCGCA TATTCGCA TTTTAGCG ATTCCTCA * *	.CAACATGA .CAACATGA CCATATAG .CGGACCGG *	TGCTATG( TGCTATG( TGCTAAT( TGCTATT( *****	CTGACCAAAC CTGACCAAAC CTGCCGCAAC STGAGCTAAA ** **	TGTTTA CCATGTTTAG CATGTTTAG CCATGTTTAG TGTTTAG *******
EC   SF   ST   KP	ATTAAAATAG ATTAAAATAG TTGAAATCAT GACAAAA-AA *** * <b>TAAA</b>	CAGGTGTTT CAGGTGTTT AACGTGCTT ATTGTACGZ ** CA	TATCCGCA TATTCGCA TTTTAGCG ATTCCTCA * *	CAACATGA CAACATGA CCATATAG .CGGACCGG *	TGCTATGO TGCTATGO TGCTAATO TGCTATTO ****	CTGACCAAAC CTGACCAAAC CTGCCGCAAC TGAGCTAAA ** **	TGTTTA CCATGTTTAG CATGTTTAG CCATGTTTAG TGTGTTTAG ******
EC   SF   ST   KP   EC	ATTAAAATAG ATTAAAATAG TTGAAATCAT GACAAAA-AA *** * TAAA TAAATTAAA	CAGGTGTTT CAGGTGTTT AACGTGCTT ATTGTACGZ ** CA CA	TATCCGCA TATTCGCA TTTAGCG ATTCCTCA * * ATGAATAC	CAACATGA CAACATGA CCATATAG CGGACCGG * TCTCCCCTG	TGCTATGO TGCTATGO TGCTAATO TGCTATTO ***** AACATTC2	CTGACCAAAC CTGACCAAAC CTGCCGCAAC STGAGCTAAA ** **	TGTTTA CATGTTAG CATGTTAG CATGTTAG TGGTTTAG *******
EC   SF   ST   KP   EC   SF	АТТААААТАG АТТААААТАG ТТGАААТСАТ GACAAAA-AA *** * ТААА ТАААТ <b>ТААА</b> ТАААТ <b>ТААА</b>	CAGGTGTTT CAGGTGTTT AACGTGCTT ATTGTACGZ ** CA CA CA AAGAAAZ	TATCCGCA TATTCGCA TTTTAGCG ATTCCTCA * * ATGAATAC	CAACATGA CAACATGA CCATATAG CCGGACCGG * * TCTCCCTG TCTCCCTG	TGCTATGO TGCTATGO TGCTATTO TGCTATTO ***** AACATTO AACATTO	CTGACCAAAC CTGACCAAAC CTGCCGCAAC GTGAGCTAAA ** ** ATGTGACGTG ATGTGACGTG	TGTTTA CATGTTAG CATGTTAG CATGTTAG TGTTTAG TTGATTATCG TTGATTATCG
EC   SF   ST   KP   EC   SF   ST	ΑΤΤΑΑΑΑΤΑG ΑΤΤΑΑΑΑΤΑG ΤΤGΑΑΑΤCΑΤ GACAAAA-AA *** * ΤΑΑΑ ΤΑΑΑΤ <b>ΤΑΑΑ</b> ΤΑΑΑΤ <b>ΤΑΑΑ</b>	CAGGTGTT CAGGTGTT AACGTGCT ATTGTACG ** CA CAAGAAAA CAAGAAAA CAAGAAAC	TATCCGCA TATTCGCA TTTAGCG ATTCCTCA * * ATGAATAC ATGAATAC ATGAATAC	CAACATGA CAACATGA CCATATAG CGGACCGG * TCTCCCTG TCTCCCTG AACTCCTG	TGCTATGO TGCTATGO TGCTAATO TGCTATTO ***** AACATTO AACATTO AACTGTO	CTGACCAAAC CTGACCAAAC CTGCCGCAAC STGAGCTAAA ** ** ATGTGACGTG ATGTGACGTG CTGTGATGTC	TGTTTA CATGTTAG CATGTTAG CATGTTAG TTGTTAG ******* TTGATTATCG TTGATTATCG
EC   SF   ST   KP   EC   SF   ST   KP	АТТААААТАG АТТААААТАG ТТGАААТСАТ GACAAAA-AA *** * ТААА ТАААТТААА ТАААТТААА ТАААТТААА	CAGGTGTT CAGGTGTT AACGTGCT ATTGTACG ** CA CAAAGAAAA CAAAGAAAA CAAAGAACC CAAAGAACC	TATCCGCA TATTCGCA TTTTAGCG ATTCCTCA * * ATGAATAC ATGAATAC ATGAATAC	CAACATGA CAACATGA CCATATAG CGGACCGG * * TCTCCCCTG ACTCCTG CACTCCTG	TGCTATGO TGCTATGO TGCTATTO TGCTATTO ***** AACATTO AACATTO AACTGTO ACTTCTO	CTGACCAAAC CTGACCAAAC CTGCCGCAAC STGAGCTAAA ** ** ATGTGACGTG ATGTGACGTG CTGTGATGTG TGTGATGTG	TGTTTA CATGTTAG CATGTTAG CATGTTAG TGTTAG TTGATTACG TTGATTATCG TTGATTATCG

Fig. 4. Conservation of NadR-sites upstream of nadB. Notation as in Fig. 3.

Unexpectedly, NadR-sites upstream of other regulon members are not well conserved in genomes other than S. typhi and E. coli.

The NadR-site upstream of nadB is conserved in *E. coli*, *Sh. flexneri*, *S. typhi*, and *K. pneumoniae* (Fig. 4).

The corresponding regions of other genomes are not conserved, nor they contain candidate NadR-sites.

The situation with pncB is somewhat more interesting (Fig. 5a).

The site is conserved in *E. coli, Sh. flexneri* and *S. typhi.* The corresponding region in *K. pneumoniae* and *S. marcescens* is not conserved, although there are two conservation islands on both sides. Thus the NadR sites were destroyed in these genomes. New candidate sites appeared instead and these new sites do not seem to originate from local duplications. Indeed, there is no sequence conservation around "old" and "new" NadR-sites (Fig. 5b).

No sites were found in the remaining genomes.

In an attempt to find new candidate members of the NadR regulon, we identified candidate sites and considered all genes with candidate sites in at least four genomes. Unexpectedly, one of such genes was nadR itself, that had a strong candidate site in *E. carotovora, S. marcescens, Y. pestis* and *Y. enterocolitica.* The alignment of the upstream regions is shown in Fig. 6.

The "four-genome" condition holds in two more cases: two genes ynfL and ynfM transcribed in opposite directions, and rpsP.

The gene ynfl encodes a putative regulator from the LysR family, whereas ynfM encodes a putative transporter. We identified ynfLM ortgologs in *Pseudomonas* spp.

EC	GAGTCTGGTG-	TTCAGTO	CTAT	TCCTGT	T	-GCGTA	AATCG	-CGCT2	ATGCA
SF	GAGTCTGGTG-	TTCAGTO	CTAT	TCCTGT	T	-GCGTA	AATCG	-CGCT2	ATGCA
KP	AAGTGTCGT	CCCAGTO	CTAT	TCCTGT	r	-GTGTC	AATCG	-CGCT2	ATGCA
ST	CACTTTCCCG-	CTATGCO	CCC-AT	CACTGC	CCCAAAG	CATGGT	AGCAG	-CGCA	GTAGA
SM	GAGCGGCAAGC	GATCGGGT	CAGCGT	GCATAC	CGAAGCC	GGCTTT	ATCTGAI	TCGCT	GTTTA
	*	* *	* *	*			* *	* * *	* *
EC	GAATCTTCATC	CTTTTCAGO	FTACAA	ACGCCT	TTATTGC	TACATT	-TTTATA	ACATA	CAC
SF	GAATCTTCATC	CTTTTCAGO	TACAA	ACGCCT	TTATTGC	TACATT	-TTTATA	ACATA	CAC
KP	GAATCTTCATC	CTTTTCAAC	GTGAA	ACACGG.	AAATCGC	TACATT	-TTGTTA	ACACT	CGCGG
ST	AATCCTTAAA-	TTCAAGO	GGTTA	GCAGTC	GCATCGC	TACATT	-TTTATA	ACATG	GGG
SM	AAATAA <b>T</b> T <b>AAC</b>	<b>A</b> TTATAA1	TTTTA	TGACTA	ATTAGGC	TAAGTC	ATTCACC	TTACA	GGCAT
	* * *	** *	*		* *	** *	* *	*	
EC	CGCGTAATGCC	CATCGACCA	AGAAAG	GTGGCA	TATGGTG	TGATCG	GGGTTCA	ATAAA	ΓT – – –
SF	CGCGTAATGCC	CATCGACCA	AGAAAG	GTGGCA	TATGGTG	TGATCG	GGGTTCA	ATAAA	TT
KP	CACGAAATGCC	CCTCGACCO	CGACGC	AAAGCT'	IGTGGTG	TGATCC	A <b>TGTT</b> C <b>A</b>	ATATA	T <b>TAAA</b>
ST	CACGAAATGCC	GCTCGACCO	TAAAG	ACAGCT	TATGGTG	TGATCG	GGGTTCA	ATAAA	rc
SM	ATCTGGCTTTT	TTTTTCTCC	CCCGTC	GCCGC-	CAGGCCG	TCATAA	AGGCACO	TTTAA	rc
	* *	* *		* *	* *	* **	* *	* *	*
EC							-GCGAAA	CA	
SF							-GCGAAA	CA	
KP	CTAGGCCTCGC	CAAATGACO	GTCAG	CGTCAC	CATTGCT	CGCCAT	CGCGGGA	CAGAG	FCGGG
ST							-GCTAAA	CA	
SM	-CTCGACCCGC	CTTTGGTGA	ATTTAT	GGTGTG.	ATGCAGC	TTCAAT	AACAGGA	TA	
							* *	*	
						_			
				~~~~~	TGTTT	A	TAAACA		~~~~
EC	AGGTA	ATACTCCAC	GCAG'I''I'	CCTGAA	GATGTTT	ATTGTA	CTAAAC	CTCCT	G'I'AC-
SF	AGGTA	ATACTCCAC	GCAG'I''I'	CCTGAA	GATGTTT	A'I'I'G'I'A	TAAAC	CTCCT	J'I'AC-
KP	TAATAAAGGTA	ATACTCCG	CTCCA	TTTTCC	GCGT'TGG	TTTCGA	IGGAACO	CTCCA	GTGA-
ST	AGG'I'A	ATACTCCAC	GCGG'I''I'	TTCTTA	GT <b>TGTTT</b>	ATTTGTA	CTAAACA	CTCCCC	G'I'GA-
SM	ATCCGGGTA	ATACTCCAC	CCCCCA	CTTTTA	TGATTAT	CCGGAT	TTGGACA	CGCGC	CTGAC
	****	******	*		*		**	* *	×
ROL	C) CO			0 7 0 7 7 m	тааатта	maamam		maamm	
EC	GAGGACGCTAC	TGCGCACC			TCGCTTC	TCCTGT			CTGG
	GAGGACGCTAC	TGCGCACC				TCCTGT:			
CL	CACCACCCAAC	TACCCCACC				TCCTGL	HCTGCAC		CITG
ST.	CACCARCOCAAC		- ATGA			CCCCGTGT"			CTGG
SPI	SAGGATGCIG	*** *	- <b>A</b> TGA	* ****	* ****	** *	* *	** **	*** *
						^			

Fig. 5a. Conservation of "old" NadR-sites upstream of pncB. Notation as in Fig. 3.

TGTTTA	TAAACA

EC	TACTCCAGCAGTTCCTGAAGA <b>TGTTTA</b> TTGTAC <b>TAAAC</b> GCTCCTGTAC-GAGGACGCTACTGCGCACCT <b>ATG</b>
SF	TACTCCAGCAGTTCCTGAAGA <b>TGTTTA</b> TTGTAC <b>TAAAC</b> GCTCCTGTAC-GAGGACGCTACTGCGCACCT <b>ATG</b>
ST	TACTCCAGCGGTTTTCTTAGT <b>TGTTTA</b> TTGTAC <b>TAAACA</b> CTCCCGTGA-GAGGACGCAACAGCGCACCT <b>ATG</b>
SM	AGCCGGCTTTATCTGATTCGC <b>TGTTTA</b> AAATAA <b>T</b> T <b>AACA</b> TTATAATTTTTATGACTAATTAGGCTAAGTCAT
KP	CAAAGCTTGTGGTGTGATCCA <b>TGTTCA</b> ATATAT <b>TAAACT</b> AGGCC-TCGCAAATGACCGTCAGCGTCACCA
	**** * ** * *

Fig. 5b. Alignment of "new" NadR-sites upstream of pncB. Notation as in Fig. 3.

#### 1014 A. V. Gerasimova & M. S. Gelfand

		TGTTTA	TAAACA
YE	-TTTAGTCGTGTTCATCGCGCAGCATACTGTGTTATTT	TGTTTAGT	gcgc <b>taaaca</b> A
YP	-TTTAATAGTGCCAATCCCGCAGCATACTGTGTTATTT	TGTTTAGT.	acgc <b>taaaca</b> a
SM	TTTTATTAATGTCGCTGACGGCGAGCAAGATATG-CTATT	TGTTTAGC.	atat <b>taaac</b> g <mark>g</mark>
ERW	GAGTAGCGGACGTTGCCTTATCTTGTG-GTATT	T <b>GTTTA</b> GT.	AGGC <b>TAAACA</b> A
	* * * * * * * * * * * *	******	* * * * *
YE	GGAGGCGGT <b>ATG</b> CTGCAGTTCGACTATCTCAAAACAGC	CGATTAAGC.	AAAAAGGCT
YP	GGAGACCGC <b>ATG</b> CTGCAGTTCGACTATCTCAAAACGGC	CAATTAAGC.	AAAAAGGCT
SM	AGGTGGCCATGCCGCAATTTGATTACCTGAAGACGTC	CCATCAAGC.	AAAAGGGTT
ERW	GAGTGTTGCACC <b>ATG</b> TCATCATTTGATTACCTGAAATCCGC	CTATCCGCC.	AGAAGGGTT
	*** ** ** ** ** **	* * * *	* ** ** *

Fig. 6. Alignment of regions upstream of *nadR*. Notation as in Fig. 3.

	TG	FTTA		TAAAC	A
EC	CTTATACATAGGGTAGGAAAATCGA-ATTGTTC <b>TG</b>	<b>T</b> C <b>TA</b> A'	FATAT	<b>TAA</b> TA	AT-CTC
SF	CTTATACATAGGGTAGGAAAATCGA-ATTGTTC <b>TG</b>	<b>P</b> C <b>TA</b> A'	FATAT	TAATA.	AT-CTC
KP	-GCTCACATTTTTAGGGTATGAAAATGTA-AATATTC <b>TG</b>	rct <b>a</b> a'	FATAT	<b>TAA</b> TA	AT-CTC
ST	-ACCGACATGTAAAGCATAGAAAAAGCAA-AATATTC <b>TG</b>	rct <b>a</b> a'	FATAT	TAATT	GT-CTC
SM	-GCAGATAACAAAATGATAGGGAGTGGCG-AATTTTT <b>TG</b>	rct <b>a</b> a'	FATAT	<b>taa</b> ta	ATTCAA
YE	-TGTAATAATAGGATCATAGAAATAGCAG-AGTTTTT <b>TG</b>	<b>F</b> C <b>TA</b> A'	FATAT	TAATT.	AT-TCA
YP	-CAGAACATTTTAATCATAGAAATAGTTT-GTTTTTT <b>TG</b>	rct <b>a</b> a'	FATAT	TAATC.	AT-TGC
ERW	-AACAATAAGCCGATCATAGAAGAGTGAT-ATTATTT <b>TG</b>	rat <b>a</b> a'	FATAT	<b>taa</b> ta	AT-CAT
PHL	TTATGAAGATCAAGCATATGAATTGCAA-AATATTT <b>TG</b>	rct <b>a</b> a'	FATAT	TAATC.	AT-TAA
PF	$GGCAATGAAA-AAATCATATAGCTGGCTA-ATGTTTC\mathbf{T}AAAAAAAAAA$	<b>T</b> CC <b>A</b> A'	FATAT	<b>T</b> GTT <b>C</b>	GA-CCT
PSY	GGCAATGAAA-AAATCATATAGCTCGCTA-ATCATTCCA	<b>T</b> CC <b>A</b> A'	FATAT	TGTTC	GA-CCT
PP	CGCAATGAAA-AAAGCATATAGCTGGCTA-ACGATTAGA	<b>T</b> CC <b>A</b> A'	FATAT	TGTTC	GA-CCT
AV	GATGCCGA-CCAGCATAGGGGGGGGGGGGGGATATTCCCGGT	<b>T</b> CC <b>A</b> A'	FATAT	TGTTC	GA-CTG
BPA	-CCGCCTGGCCACAGTAGACTTCCGGC-CGCCATT <b>TG</b>	<b>F</b> CC <b>A</b> A'	FATAT	GAAAC	CTGCAC
	* *	* **	* * * * *		
EC	AAATAAGATGTTTTAAAT <b>ATG</b> AATA	TTGAA	CTTCG	TCATC	TGCGTT
SF	AAATAAGATGTTTTAAAT <b>ATG</b> AATA	TTGAA	CTTCG	TCATC	TGCGTT
KP	AAATAAGACGTTTTTAAAT <b>ATG</b> AATA	rcgag	FTGCG	TCATT	TGCGCT
ST	AAATAAGACGTTAAAAAT <b>ATG</b> AATA	TCGAA	FTACG	TCATT	TACGTT
SM	-ATTAAGACGTTTTTAAATATGAATA	rcgag	CTGAG	ACACC	TGCGTT
YE	TAATAAGACTTTAAAAATATCACTGGAGTTGG <b>ATG</b> AGTA	TTGAA	CTCAG	GCATT	TACGTT
YP	AAATAATACGTTTAAAATATCA <b>ATG</b> AGTA	TTGAA	CTAAG	ACATT	TGCGGT
ERW	GATTGATACGTTTAACATATGAATA	TAGAA	CTGCG	TCACC	TTCGCT
PHL	TAATAATATGTATTAGATCTCAAAGGTGATT- <b>ATG</b> GCTA	ICGAA'	FTACG	GCATT	TACGCT
PF	GTTTGATAGCTTTTACGACCTAATGGGGTGCCCA	<b>rg</b> gaa'	FTGCG	TCATC	TGCGCT
PSY	GTTTGATAGGTAAAACGACTTAATGGAGGCCGC <b>A</b>	<b>rg</b> gaa'	FTGCG	TCATC	TGCGTT
PP	$\texttt{ATTTGAGATGTTTTACGACTTGATTGGAGCG}{\texttt{GCTC}}{\textbf{A}}$	<b>rg</b> gaa	CTGCG	TCATC	TGCGTT
AV	ATTTGATATGTTCTACGAATCAATGGGGGCTGTCA	<b>rg</b> gag	CTGCG	TCATC	TGCGCT
BPA	СААТGАТАТТТТААААААА	<b>rg</b> gaa	CTGCG	TCACC	TGCGCT
	* * * * *	* *	* *	* *	* ** *

Fig. 7. Alignment of regions upstream of *ynfL*. Notation as in Fig. 3.
Notation: "PF" — Pseudomonas fluorescens CHA0, "PSY" — Pseudomonas syringae, "PP" — Pseudomonas putida, "AV" — Azotobacter vinelandii, "BPA" — Bordetella parapertussis.

and in *Bordetella parapertussis* and constructed multiple alignment of the intergenic region in all considered genomes (Fig. 7).

The conserved region coincides with the spacer of the candidate NadR binding site. On the other hand, there is no NadR regulator in *B. parapertussis* and in

EC	CAGCAGAGTTAGCAACTGTTGATTGCAATTCCCGCAAAAATGAGTAAAAT
SF	GCGACAGAGTTATTAACTGCTGATTGCATTTCTCCAGAAATCAGTAAAAT
ST	${\tt CCCTTGTTTCGTGTTCAGTAGCAAGATGCTGATTGCATTTTCCCCAGAAATCAGTAAAAT}$
YP	TG-TGCGAACAGGAATCTACGCTTTAGATTGCTTTTTTCGCCAAAATGAGTAAAAT
YE	TTTTGCGCCAAAATGAGTAAAAT
ERW	AC-TCTTGGCGAGATTATACGCTTTAGATTGCTTTTTGCGCCAAAATGAGTAAAAT
PHL	AG-TCTCGAAGAATATCCACTTTGGATTGCTTTTTGCGCCAAAGTGAGTAAACT
'	** * ** * ***** *
	TGTTTA
EC	TTTCGGGCTTTTAATATGACACCGGACTCCGTTCCTCGATGGGGTCCGGT
SF	TTTCGGGCTTTTAATATGACACCGGACTCCGTTCCTCGATGGGGTCCGGT
ST	TTTCGGGCTTTTAATATGACGCCGGGCTCCGTTCCTCGATGAGGCCCGGT <b>TGTTT</b> TAT
YP	TTTCGGGCTTTTTATATTGCA-ACTGGACCCCGTTCCCCGATGGGGTCCAGT <b>TGTTT</b> TAT
YE	TTTCGGGCTTTTTATATTGCA-ACTGGACCCCGTTCCCCGATGGGGTCCAGT <b>TGTTT</b> TAT
ERW	TTTCGGGCTTTTTTATATGACATACTGGGCTCCGTTCCTCGATGGGGCCCCGGT <b>TGTT</b> TAT
PHL	TTTCGGGCTTTTTTATCTGACA-ACCGGACTTCGTTCCTCGATGAAGTCTGGT <b>TGTT</b> TGT
1	******* ** * * * * * ******************
	ТАААСА
EC	TCACACAAGAGAGATGTTATGGTAACTATTCGTTTAGCACGTCACGGCGCTAAAAAGCGTC
SE	ТСАСАСААСА ССАСААСА ССАСААСА ССАССАСА ССАССА
ST	ͲϹϪϹϪϹϪϪ;;
VP	
VE	
EBM	
PHT.	
	* ** * *******************************

Fig. 8. Alignment of regions upstream of rpsP. Notation as in Fig. 3.

*Pseudomonas* spp., and thus this region cannot be a NadR-site. Since the arrangement where a binding site occurs between a divergently transcribed regulator gene and a regulated operon is very common, we conclude that the conserved region is the YnfL binding site. However, it is a very tentative prediction, requiring an experimental verification.

The gene rpsP encodes small ribosomal subunit protein S16. The nucleotide sequence of the rpsP upstream regions is uniformly conserved (Fig. 8).

This fact and the function of RpsP makes it unlikely that the observed site is functional.

### 4. Conclusions

This study demonstrated that even a very simple regulon covering an essential methabolic pathway could be different in closely related genomes. Not only the set of regulated genes can vary, but the autoregulation of the nadR gene by NadR, predicted here for the first time, is a feature of several, but not all genomes.

One of the possible explanations could be that the NadR regulon itself is rather young, as it exists in only one family of gamma-proteobacteria. However, the same behavior was observed for a number of other regulons, in particular Lrp,<sup>46,47</sup>

 $\rm FruR,^{46}$  KdgR.  $^{25}$  More sequenced genomes are needed to elucidate the exact history of the NadR regulon.

## Acknowledgments

We are grateful to Andrei Osterman, Dmitry Rodionov, Dmitry Ravcheev and Gavin H. Thomas for useful discussions. This study was partially supported by grants from the Howard Hughes Medical Institute (55000309) and Russian Academic of Sciences (Programs "Molecular and Cellular Biology" and "Origin and Evolution of the Biosphere").

## References

- Gelfand MS, Novichkov PS, Novichkova ES, Mironov AA, Comparative analysis of regulatory patterns in bacterial genomes, *Brief Bioinform* 1:357–371, 2000.
- Stojanovic N, Florea L, Riemer C, Gumucio D, Slightom J, Goodman M, Miller W, Hardison R, Comparison of five methods for finding conserved sequences in multiple alignments of gene regulatory regions, *Nucleic Acids Res* 27:3899–3910, 1999.
- Gelfand MS, Computational identification of regulatory sites in DNA sequences, in Frasconi P, Shamir R (eds.), Artificial Intelligence and Heuristic Methods in Bioinformatics, San-Miniato, Italy, IOS Press, pp. 148–172, 2003.
- Mironov AA, Koonin EV, Roytberg MA, Gelfand MS, Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes, *Nucleic Acids Res* 27:2981–2989, 1999.
- Gelfand MS, Koonin EV, Mironov AA, Prediction of transcription regulatory sites in Archaea by a comparative genomic approach, *Nucleic Acids Res* 28:695–705, 2000.
- 6. Tan K, Moreno-Hagelsieb G, Collado-Vides J, Stormo GD, A comparative genomics approach to prediction of new members of regulons, *Genome Res* 11:566–584, 2001.
- Makarova KS, Mironov AA, Gelfand MS, Conservation of the binding site for the arginine repressor in all bacterial lineages, *Genome Biol* 2:RESEARCH0013, 2001.
- Laikova ON, Mironov AA, Gelfand MS, Computational analysis of the transcriptional regulation of pentose utilization systems in the gamma subdivision of Proteobacteria, *FEMS Microbiol Lett* 205:315–322, 2001.
- Panina EM, Mironov AA, Gelfand MS, Comparative analysis of FUR regulons in gamma-proteobacteria, *Nucleic Acids Res* 29:5195–5206, 2001.
- 10. Permina EA, Mironov AA, Gelfand MS, Damage-repair error-prone polymerases of eubacteria: association with mobile genome elements, *Gene* **293**:133–140, 2002.
- 11. Rodionov DA, Mironov AA, Gelfand MS, Conservation of the biotin regulon and the BirA regulatory signal in Eubacteria and Archaea, *Genome Res* **12**:1507–1516, 2002.
- Panina EM, Mironov AA, Gelfand MS, Comparative genomics of bacterial zinc regulons: enhanced ion transport, pathogenesis, and rearrangement of ribosomal proteins, *Proc Natl Acad Sci U S A* 100:9912–9917, 2003.
- Liu J, Tan K, Stormo GD, Computational identification of the Spo0A-phosphate regular that is essential for the cellular differentiation and development in Grampositive spore-forming bacteria, *Nucleic Acids Res* 31:6891–6903, 2003.
- 14. Permina EA, Gelfand MS, Heat shock (sigma32 and HrcA/CIRCE) regulons in beta-, gamma- and epsilon-proteobacteria, J Mol Microbiol Biotechnol 6:174–181, 2003.
- Erill I, Escribano M, Campoy S, Barbe J, In silico analysis reveals substantial variability in the gene contents of the gamma proteobacteria LexA-regulon, *Bioinformatics* 19:2225–2236, 2003.

- Manson McGuire A, Church GM, Predicting regulons and their cis-regulatory motifs by comparative genomics, *Nucleic Acids Res* 28:4523–4530, 2000.
- McGuire AM, Hughes JD, Church GM, Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes, *Genome Res* 10:744–757, 2000.
- McCue L, Thompson W, Carmack C, Ryan MP, Liu JS, Derbyshire V, Lawrence CE, Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes, *Nucleic Acids Res* 29:774–782, 2001.
- Rajewsky N, Socci ND, Zapotocky M, Siggia ED, The evolution of DNA regulatory regions for proteo-gamma bacteria by interspecies comparisons, *Genome Res* 12: 298–308, 2002.
- Rodionov DA, Dubchak I, Arkin A, Alm E, Gelfand MS, Reconstruction of regulatory and metabolic pathways in metal-reducing delta-proteobacteria, *Genome Biol* 5:R90, 2004.
- Terai G, Takagi T, Nakai K, Prediction of co-regulated genes in Bacillus subtilis on the basis of upstream elements conserved across three closely related species, *Genome Biol* 2:RESEARCH0048, 2001.
- Alkema WB, Lenhard B, Wasserman WW, Regulog analysis: detection of conserved regulatory networks across bacteria: application to Staphylococcus aureus, *Genome Res* 14:1362–1373, 2004.
- Rodionov DA, Mironov AA, Rakhmaninova AB, Gelfand MS, Transcriptional regulation of transport and utilization systems for hexuronides, hexuronates and hexonates in gamma purple bacteria, *Mol Microbiol* 38:673–683, 2000.
- Ramirez-Santos J, Collado-Vides J, Garcia-Varela M, Gomez-Eichelmann MC, Conserved regulatory elements of the promoter sequence of the gene rpoH of enteric bacteria, *Nucleic Acids Res* 29:380–386, 2001.
- Rodionov DA, Gelfand MS, Hugouvieux-Cotte-Pattat N, Comparative genomics of the KdgR regulon in Erwinia chrysanthemi 3937 and other gamma-proteobacteria, *Microbiology* 150:3571–3590, 2004.
- Yellaboina S, Ranjan S, Chakhaiyar P, Hasnain SE, Ranjan A, Prediction of DtxR regulon: identification of binding sites and operons controlled by Diphtheria toxin repressor in Corynebacterium diphtheriae, *BMC Microbiol* 4, 2004.
- Erill I, Jara M, Salvador N, Escribano M, Campoy S, Barbe J, Differences in LexA regulon structure among Proteobacteria through in vivo assisted comparative genomics, *Nucleic Acids Res* 32:6617–6626, 2004.
- Florea L, McClelland M, Riemer C, Schwartz S, Miller W, EnteriX 2003: visualization tools for genome alignments of Enterobacteriaceae, *Nucleic Acids Res* 31:3527–3532, 2003.
- Yellaboina S, Seshadri J, Kumar MS, Ranjan A, PredictRegulon: a web server for the prediction of the regulatory protein binding sites and operons in prokaryote genomes, *Nucleic Acids Res* 32(Web Server issue):W318–320, 2004.
- Begley TP, Kinsland C, Mehl RA, Osterman A, Dorrestein P, The biosynthesis of nicotinamide adenine dinucleotides in bacteria, *Vitam Horm* 61:103–119, 2001.
- Kurnasov OV, Polanuyer BM, Ananta S, Sloutsky R, Tam A, Gerdes SY, Osterman AL, Ribosylnicotinamide kinase domain of NadR protein: identification and implications in NAD biosynthesis, J Bacteriol 184:6906–6917, 2003.
- Penfound T, Foster JW, NAD-dependent DNA-binding activity of the bifunctional NadR regulator of Salmonella typhimurium, J Bacteriol 181:648–655, 1999.
- 33. Foster JW, Park YK, Penfound T, Fenger T, Spector MP, Regulation of NAD metabolism in Salmonella typhimurium: molecular sequence analysis of the bifunctional nadR regulator and the nadA-pnuC operon, J Bacteriol 172:4187–4196, 1990.

- 1018 A. V. Gerasimova & M. S. Gelfand
- 34. Blattner FR *et al.*, The complete genome sequence of Escherichia coli K-12, *Science* **277**:1453–1474, 1997.
- Wei J et al., Complete genome sequence and comparative genomics of Shigella flexneri serotype 2a strain 2457T, Infect Immun 71:2775–2786, 2003.
- Parkhill J et al., Complete genome sequence of a multiple drug resistant Salmonella enterica serovar Typhi CT18, Nature 413:848–852, 2001.
- Bell KS et al., Genome sequence of the enterobacterial phytopathogen Erwinia carotovora subsp. atroseptica and characterization of virulence factors, Proc Natl Acad Sci USA 101: 11105–11110, 2004.
- Parkhill J et al., Genome sequence of Yersinia pestis, the causative agent of plague, Nature 413:523–527, 2001.
- Duchaud E et al., The genome sequence of the entomopathogenic bacterium Photorhabdus luminescens, Nat Biotechnol 21:1307–1313, 2003.
- Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL, GenBank, Nucleic Acids Res 31:23–27, 2003.
- Schneider TD, Stephens RM, Sequence logos: a new way to display consensus sequences, Nucleic Acids Res 18: 6097–6100, 1990.
- Tatusov RL, Galperin MY, Natale DA, Koonin EV, The COG database: a tool for genome-scale analysis of protein functions and evolution, *Nucleic Acids Res* 28:33–36, 2000.
- Felsenstein J, Evolutionary trees from DNA sequences: a maximum likelihood approach, J Mol Evo 17:368–376, 1981.
- 44. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG, The CLUSTAL\_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools, *Nucleic Acids Res* 24:4876–4882, 1997.
- Mironov AA, Vinokurova NP, Gel'fand MS, Software for analyzing bacterial genomes, Mol Biol (Mosk) 34:253–262, 2000.
- Friedberg D, Midkiff M, Calvo JM, Global versus local regulatory roles for Lrp-related proteins: Haemophilus influenzae as a case study, J Bacteriol 183:4004–4011, 2001.
- Gelfand MS, Laikova ON, in Michael YG and Eugene VKC (eds.), Frontiers in Computational Genomics, Academic Press, Wymondham, UK, pp. 203–204, 2003.

**Anna V. Gerasimova** is a graduate student in the Laboratory of Bioinformatics, State Scientific Center GosNIIGenetika, Moscow, Russia. She obtained her M.S. in Biophysics from Moscow Engineering Physics Institute (Technical University) in 2002. Her research interests are in comparative genomics and molecular evolution.

Mikhail S. Gelfand is the Head of the Research and Training Center in Bioinformatics of the Institute for Information Transmission Problems, RAS in Moscow, Russia and a Professor at the Department of Bioengineering and Bioinformatics of the Moscow State University. He graduated from the Department of Mathematics of the Moscow State University, received his Ph.D. (Math.) degree from the Institute of Theoretical and Experimental Biophysics, RAS (Pushchino), and the Doctor of Sciences degree from the State Research Institute for Genetics and Selection of Industrial Microorganisms (Moscow). He is a member of editorial boards of several journals, in particular, "PLoS Biology", "Bioinformatics", "BMC Bioinformatics", "Journal of Bioinformatics and Computational Biology" and "Journal of Computational Biology". He received the A. A. Baev prize (1999) from the Russian State "Human Genome" Council, and "The Best Scientist of the Russian Academy of Sciences" award (2004). His research interests include comparative genomics, metabolic reconstruction and modeling, evolution of metabolic pathways and regulatory systems, function and evolution of alternative splicing, functional annotation of genes and regulatory signals.