ELSEVIER

FEMS Microbiology Letters 205 (2001) 305–314

FEMS MICROBIOLOGY Letters

www.fems-microbiology.org

Transcriptional regulation of pentose utilisation systems in the *Bacillus/Clostridium* group of bacteria

Dmitry A. Rodionov [a,*], Andrey A. Mironov [b], Mikhail S. Gelfand [b]

[a] *State Scientific Center GosNIIGenetika, Moscow 113545, Russia*
[b] *Integrated Genomics, Moscow, P.O. Box 348, Moscow 117333, Russia*

Received 14 August 2001; received in revised form 28 September 2001; accepted 22 October 2001

First published online 15 November 2001

Abstract

In *Bacillus subtilis*, utilisation of xylose, arabinose and ribose is controlled by the transcriptional factors XylR, AraR and RbsR, respectively. Here we apply the comparative approach to the analysis of these regulons in the *Bacillus/Clostridium* group. Evolutionary variability of operon structures is demonstrated and operator sites for the main transcription factors are predicted. The consensus sequences for the XylR and RbsR binding sites vary in different subgroups of genomes. The functional coupling of gene clusters and the conservation of regulatory sites allow for detection of non-orthologous gene displacement of ribulose kinase in *Enterococcus faecium* and *Clostridium acetobutylicum*. Moreover, candidate catabolite responsive elements found upstream of most pentose-utilising genes suggest CcpA-mediated catabolite repression. © 2001 Published by Elsevier Science B.V. on behalf of the Federation of European Microbiological Societies.

Keywords: Computer analysis; Transcriptional regulation; AraR; XylR; RbsR; CcpA; Ribulokinase

1. Introduction

Identification of regulatory sites using genome and proteome comparisons in complete bacterial genomes is an important step in genome annotation. The comparative approach based on the assumption that regulons (sets of co-regulated genes) are conserved in related genomes allows one to make reliable predictions of regulatory sites if several related genomes are available. This allows one to identify known regulons in poorly characterised organisms, predict new members of regulons, and even describe regulons *de novo*. Previously, we have applied the comparative approach to the analysis of the GntR, UxuR/ExuR and KdgR regulons that are involved in the sugar metabolism in gamma purple bacteria [1], the purine, arginine and aromatic amino acids regulons [2], the heat shock, SOS and multiple drug resistance regulons of eubacteria [3], as well as several archaeal regulons [4], reviewed in [5–6]. Other groups have applied the comparative approach to large-scale analysis of regulation in gamma-proteobacteria [7–10]. Clearly, this approach can be applied only if orthologous regulators are present in the studied genomes.

Bacillus subtilis and other Gram-positive bacteria from the *Bacillus/Clostridium* group can use various carbohydrates as a single source of carbon and energy [11]. Using extracellular carbohydrases, bacilli degrade several polysaccharides that are widely distributed in plant cell walls. Thus produced oligo-, di- or monosaccharides are transported into the cell, phosphorylated and subsequently catabolised via glycolysis or the pentose phosphate pathway. The catabolised monosaccharides can be subdivided into numerous hexoses (including hexitols and hexuronic acids) and several pentoses, namely arabinose, xylose and ribose. Catabolic enzymes are usually synthesised only when their substrate is present and the preferred carbon and energy sources are absent. The induction of the *B. subtilis* pentose catabolic operons, as well as global carbone catabolite repression, is mediated by several transcriptional repressors. The utilisation of arabinose, xylose and ribose is controlled by AraR, XylR and RbsR, respectively. The catabolite repressor protein CcpA represses transcription of catabolic genes by binding to a palindromic sequence called CRE (catabolite responsive element).

The utilisation of arabinose in *B. subtilis* is controlled by the transcription factor AraR which belongs to the LacI/

* Corresponding author. Fax: +7-095-315-0501.
E-mail address: rodionov@genetika.ru (D.A. Rodionov).

PII: S0378-1097(01)00494-3

GalR family of repressors. In the absence of inducer, AraR binds to operator sites in the promoter regions of the *araABDLMNPQ-abfA*, *araE* and *araR* operons [12]. The *araABD* genes encode three intracellular enzymes for arabinose catabolism, arabinose isomerase, ribulokinase and ribulose-5-phosphate epimerase, respectively. The function of *araL* and *araM* is unknown. The *araE* gene encodes a proton symporter involved in the transport of arabinose into the cell. AraE displays a broad substrate specificity for various sugars, namely arabinose, xylose and galactose, but its only known inducer is arabinose [13]. The putative products of *araN*, *araP* and *araQ* are homologous to components of the binding protein-dependent transport systems involved in the high-affinity transport of malto-oligosaccharides and multiple sugars. The last gene of the *ara* operon, *abfA*, encodes α-arabinofuranosidase that releases arabinose monomers from oligoarabinosides. The fact that *araLMNPQ* and *abfA* are not essential for the utilisation of arabinose suggests that arabinose oligomers are a native substrate of AraNPQ [14].

The ribose transport and the subsequent phosphorylation in *B. subtilis* are mediated by the ATP-binding cassette transport system (ABC system) *rbsABCD* and ribokinase encoded by *rbsK* [15]. These genes form an operon with the gene *rbsR* encoding the repressor of the *rbs* operon. The binding signal of RbsR is unknown. The presence of the CRE sequence in the regulatory region of the *rbs* operon demonstrates that this operon is also controlled by the catabolite repressor CcpA. CcpA and RbsR are homologous (31% identity) and belong to the LacI/GalR family of transcriptional regulators.

The utilisation of xylose in the *Bacillus* genomes is negatively controlled by the XylR repressor [16–18]. In the absence of xylose, XylR binds to the operator sites upstream of the *xylAB* operon. In addition, in *B. subtilis* XylR controls the *xynCB* operon which encodes β-xyloside permease and β-xylosidase [19]. Finally, the expression of the *Bacillus stearothermophilus* gene *xyn1* encoding extracellular xylanase T-6 is induced on the transcriptional level by xylose [20].

The transcriptional repressor CcpA mediates the catabolite repression in *B. subtilis* and related bacteria by binding to CREs. In *B. subtilis*, the genes for the utilisation of arabinose and xylose are under catabolite repression [12,21,22]. The xylose operon of *Bacillus megaterium* is also negatively regulated by CcpA [23].

Here, we apply the comparative approach to the analysis of the transcriptional regulation of catabolic and transport genes for the utilisation of pentose sugars in various genomes of the *Bacillus/Clostridium* group of bacteria. We predict new members of the arabinose and xylose regulons and describe the differences between these regulons in the analysed genomes. We also identify candidate CRE boxes that are consistently observed upstream of the pentose catabolic operons.

2. Materials and methods

The comparative analysis involved all available genomes from the *Bacillus/Clostridium* group containing genes of the arabinose, xylose or ribose pathways. The complete genome sequences of *B. subtilis*, *Bacillus halodurans* and *Lactococcus lactis*, as well as partial sequences of *B. stearothermophilus*, *B. megaterium*, *Bacillus licheniformis*, *Bacillus* sp., *Clostridium acetobutylicum*, *Staphylococcus xylosus*, *Thermoanaerobacter ethanolicus*, *Lactococcus brevis*, *Lactococcus pentosus* and *Lactobacillus sakei* were downloaded from the GenBank database [24]. Preliminary sequence data of *Bacillus anthracis*, *Staphylococcus aureus* and the genomes of *Enterococcus faecalis*, *Enterococcus faecium*, *B. stearothermophilus* and *Clostridium difficile* were obtained from WWW sites of The Institute for Genomic Research (http://www.tigr.org), DOE Joint Genome Institute (http://www.jgi.doe.gov), University of Oklahoma's Advanced Center for Genome Technology (http://www.genome.ou.edu) and the Sanger Centre (http://www.sanger.ac.uk), respectively. The three completed genomes of streptococci have not been analysed since they do not contain the considered metabolic systems.

The existence of the regulatory gene encoding the corresponding transcription factor is a pre-requisite to the comparative analysis. In the case of known AraR and XylR regulons, the training sets consisted of upstream regions of genes known to be co-regulated. In the case of the local RbsR and XylR regulons, the training sets contained upstream regions orthologous genes from related genomes, since the number of target genes in each genome was small. However, if we observed systematic differences between predicted sites in symmetrical positions of the palindromic signal, as in the case of XylR, we split the training set into homogeneous subsets.

A simple iterative procedure implemented in the program *SignalX* was used for construction of a profile from a set of upstream gene fragments [4]. Weak palindromes are selected in each region. Each palindrome is compared to all other palindromes, and the palindromes most similar to the initial one are used to make a profile. The positional nucleotide weights in this profile are defined as:

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

where $N(b,k)$ is the count of nucleotide b in position k [2]. The candidate site score is the sum of the respective positional nucleotide weights:

$$Z(b_1 \ldots b_L) = \sum_{k=1 \ldots L} W(b_k, k)$$

where k is the length of the site. Z-score can be used to assess the significance of an individual site.

These profiles are used to scan the set of palindromes again, and the procedure is iterated until convergence. Thus a set of profiles is constructed. The quality of a profile is defined as its information content [25]:

$$I = \sum_{k=1\ldots L} \sum_{i=\mathrm{A,C,G,T}} f(i,k)\log(f(i,k)/0.25)$$

where $f(i,k)$ is the frequency of nucleotide i in position k of sites generating the profile. The best profile is used as the recognition rule.

Each genome was scanned with the profile, and genes with candidate regulatory sites in the upstream regions (in positions −325 to +25 relative to the translation start) were selected. The threshold for the site search was defined as the lowest score observed in the training set.

Protein alignments were made using the Smith–Waterman algorithm implemented in the *GenomeExplorer* program [26]. Orthologous proteins were defined by the best bidirectional hits criterion [27]. Distant homologues were identified using PSI-BLAST [28]. Multiple sequence alignments were constructed using CLUSTALX [29]. Phylogenetic trees were created by the maximum likelihood method implemented in PHYLIP [30] and drawn using TreeView [31]. Site recognition was performed using *GenomeExplorer*.

3. Results and discussion

3.1. The arabinose regulon

The arabinose regulons in the complete *B. halodurans* and *C. acetobutylicum* genomes and in the unfinished *B. stearothermophilus* and *E. faecium* genomes were identified using the similarity search with genes from the arabinose regulon of *B. subtilis* as queries (Fig. 1A). The AraR search profile was constructed using the training set of five known AraR-binding sites from *B. subtilis* [12] (see Section 2). The threshold was set to 4.86, the minimal score of the five sites from the training set. Several new candidate AraR-regulated genes were identified (Table 1A). Of these, only two new genes, *xsa* and *ydjK*, were predicted to belong to the AraR regulon in *B. subtilis*. The former encodes the second α-arabinofuranosidase homologous to *abfA* from the *ara* operon. The latter encodes a putative transporter from the Major Facilitator Superfamily, the Sugar Porter subfamily [32] and is homologous to various arabinose, xylose and galactose symporters. Interestingly, it has been shown recently that *B. subtilis* has a low affinity arabinose transporter in addition to *araE* and *araNPQ* [33]. Thus, the hypothetical proton symporter *ydjK*, predicted to belong to the AraR regulon, can be also involved in the transport of arabinose. In addition, the α-arabinase gene *abnA*, located immediately upstream of the *araABDLMNPQ-abfA* operon, has a weak candidate AraR site (score of 4.49). α-Arabinase is an extracellular enzyme involved in the degradation of arabinose polymers. Thus this gene can be co-regulated with the rest of the arabinose-utilising genes.

In *B. halodurans*, the predicted AraR regulon consists of the *araDBA-xsa*, *abfA-araM* and *araR* operons. Furthermore, we identified a strong candidate AraR site upstream of *BH1061*. This gene encodes a putative sugar hydrolase closely related to α-glucuronidase AguA from *Thermotoga maritima* (54% identity). The orthologue of this gene in *B. stearothermophilus* belongs to the glucuronic acid utilisation operon which is regulated by the local repressor UxuR [34]. No candidate AraR sites were observed upstream of this operon. Nevertheless, the regulation of *BH1061* by the arabinose repressor can be significant, since glucuronosyl residues appear in such natural heteropolysaccharides as arabinans and arabinoxylan. The *B. halodurans* transport system *araNPQ* has no AraR sites, but it is likely to be regulated by its own transcriptional regulator (LacI/GalR family) encoded by the upstream gene *BH0901*. In the regulatory region of *araNPQ* we found a probable operator, a 24-bp perfect palindromic sequence TTTTTCGTGTACGTACACGAAAAA. Its 10 central positions coincide with the corresponding positions of the AraR consensus site, ATTTGTACGTACAAAT. It is likely that these related transcriptional factors have similar recognition sequences.

The search for orthologues in the unfinished genome of *B. stearothermophilus* reveals only two fragments containing genes of the arabinose metabolism. The first one contains *araR, araDBA* and partially the *araGH* transport system. The second one contains a fragment of the *xsa-abfA* operon lacking the upstream region. High scoring AraR sites were detected upstream of the *araR* and *araDBA* operons.

In the more distant *C. acetobutylicum* genome, the composition and the operon structure of the AraR regulon is different. There are no orthologues of the *araB*-encoded ribulose kinase. Moreover, *araE* and *araA* are duplicated and combined into a single gene cluster with *araD, araR* and two new genes, *CAC1343* and *CAC1344*. The first of these genes is orthologous to the phosphoketolase gene *ptk* from *L. lactis* and the second one, named *araK*, is similar to several sugar kinases, but does not cluster on the phylogenetic tree with any kinases of known specificity (Fig. 2). The regulatory regions of *araDA1, araR, ptk* and *araK* each contain at least one AraR site with the score exceeding the threshold. Moreover, in *E. faecium*, the orthologue of *araK* also is predicted to belong to the AraR regulon, as it is located in the putative *araKDA* operon. We propose that the new gene *araK* belongs to the AraR regulon and encodes the missing ribulose kinase that was non-orthologously replaced in the *C. acetobutylicum* and probably *E. faecium* genomes. The *araE1* gene which is located downstream of *araR* has only a low-scoring AraR site (score of 4.49). The paralogues of the other two genes, *araE2* and *araA2,* are located immediately downstream of *araK* and

Table 1
The AraR (A), RbsR (B) and XylR (C) regulatory sites in the *Bacillus/Clostridium* group of bacteria

Genome	Gene (operon)	AC	The AraR, RbsR or XylR site			The CcpA site		
			Sequence	Pos	Score	Sequence	Pos	Score
(A) The AraR regulon								
B. subtilis	*araE/araR*	C.G.	*ATTTGTACGTACtAAT*	−106	5.83	aTGAAAaCGCTTTact	−37	4.43
			ATaaGTACGTACAAtT	−63	5.44			
			ATgTaTACGgACAAAT	−187	4.98			
	araABDLMNPQ-abfA	C.G.	*AaTTGTtCGTACAAAa*	−102	5.03	TTGAAAGCGtTTTatt	−37	4.60
			ATTaGTACGTAtcttT	−58	4.86			
	ydjK	C.G.	ATTTtTACGTACAAtT	25	5,19	aTGAAAtCGCTTcCAt	−174	4.57
	xsa	C.G.	ATacaTACGTACAAAT	−170	4.92	TTaAAAGCGCTTaCAt	−99	4.89
	abnA	C.G.	tTTTGTctGTACAAAT	−156	4.54	TTGtAAGCGCTTTCtA	−37	5.05
B. halodurans	*abfA-araM*	C.G.	ATTTGTACGTACAAAa	−197	5.56	–		
			tTaTGTACGTACAAgT	−99	4.75			
	araDBA-xsa	C.G.	AaTTGTACGTACAAgT	−70	5.22	aTGtAAGCGgTaTCgA	87	4.24
	araR	C.G.	ATTTGTACGTAtAAgT	−35	5.12	–		
	BH1061	C.G.	ATTcGTACGaACAAAT	−171	5.04	–		
B. stearothermophilus	*araR*	AF160811	AaTgaTACGgACAAAT	−68	4.86	–		
	araDBA	AF160811	AaTTGTACGTACAAta	−54	5.24	–		
C. acetobutylicum	*ptk (CAC1343)*	C.G.	ATTTaTACGTACAAAT	−101	5.70	–		
	abf2-CAC1530	C.G.	ATTTaTACGTAtAAAT	−188	5.44	–		
			ATaTGTACGTAtAtAT	−47	5.24			
	araDA/araR	C.G.	ATTTaTACGTAtcAAT	−87	5.12	–		
			ATTTaTAaGTACAAAT	−244	5.05			
	araK-araE2-araA2*	C.G.	ATagGTACGTACcAtT	−78	4.93	aTGAAAtCGtTTaaAA	−41	4.27
			AcaTGTACGTACAAAa	−47	4.75			
	araE1	C.G.	ATTcaTACGTAtAAAa	−214	4.49	agGtAAtCGCTTTCAt	−151	4.18
						aTGAAAaCGtTaTatA	−92	4.24
E. faecium	*araK*-araDA/araE*	U.G.	tTaTGTACGTACAAAa	−143	4.93	TTGtAAGCGCTactAA	−61	4.53
	araR	U.G.	ATaTGTACGTACAAAT	−30	5.73			
	abfA-…	U.G.	–			TTGAcAaCGCTTTCAA	−57	4.97
	Consensus:		ATTTGTACGTACAAAT					
(B) The RbsR regulon								
B. subtilis	*rbsRKDACB*	C.G.	*TcTATGTAAaCGGTTACATAAA*	−38	6.92	**aTGtAAaCGgTTaCAt**	−35	4.79
B. stearothermophilus	*rbsRKDACB*	U.G.	*TATgTGTAACCGGTTACAcATA*	−45	7.08	gTGtAAcCGgTTaCAc	−42	4.02
B. anthracis	*rbsRKDACB*	U.G.	*TTTcTGTAAaCGGTTACATgAA*	−18	6.55	cTGtAAaCGgTTaCAt	−15	4.45
B. halodurans	*rbsRKDACB*	C.G.	*TATcTGTAACCGGTTACAcATA*	−31	7.02	cTGtAAcCGgTTaCAc	−34	4.02
C. difficile	*rbsRKBAC*	U.G.	*TcTAgGTAACCGGTTAagTAAA*	−173	6.19	–		–
L. lactis	*rbsRKDACB*	C.G.	*TTaATGaAAaCGGaTACAatTA*	−66	4.43	ATGAAAaCGgaTaCAA	−63	4.64
	Consensus:		**TtTaTGTAAcCGgTTACAtAaA**					
L. sakei	*rbsUDKR*	AF115391	*TAGTAAAACGTTTTACTA*	−49	7.08	TgtAAAcCGtTTTCAA	−115	4.35
E. faecalis	*rbsUDK*	U.G.	*aAtTAAAACGTTTTACTA*	−62	6.50	aTGAAcGCGgTTaCAA	5	4.45
			TAcTAAAACGTTTTAaTt	−49	5.97			
S. aureus	*rbsKDU*	U.G.	*TgGTAAAACGTTTTACTA*	−43	6.79	aTGAAAaCGCaTgCAt	−87	4.31
	Consensus:		**TAgTAAAACGTTTTAcTA**					

Table 1 (*Continued*)

Genome	Gene (operon)	AC	The AraR, RbsR or XylR site			The CcpA site		
			Sequence	Pos	Score	Sequence	Pos	Score
(C) The XylR regulon								
B. subtilis	*xylAB/xylR*	C.G.	***GTTTGTTTAaaCAAcAAACtAAt***	−96	6.08	**TTGgAAGCGCaaaCAA**	33	4.67
	xynCB	C.G.	***GTTaGTTTgTTTgATcAACAAAC***	−320	6.21	**TTGAAAGCGCTTTtAt**	−100	5.01
	ygaE	C.G.	aaTTGTTTAgcAAATAAACtAAC	−94	5.62	–		
B. stearothermophilus	*xylAB*	U.G.	*cTTTGTTTATaTgATAgACAAAC*	−120	6.12	aTGAAAGCGtTaTaAt	−93	4.72
	xylR	U.G.	*aTTaGTTTATTAtATAAACtAAg*	−260	6.38	–		
	xyn1	Z29080	***GTTaGTaTATTTAATttACAAAC***	−139	5.96	AgGAAAaCcCTTTCAt	−55	4.06
B. megaterium	*xylABT/xylR*	Z71474	***GTTaGTTTATTGgATAAACAAAC***	−86	6.77	**TTGAAAGCGCaaaCAA**	33	4.90
B. licheniformis	*xylAB/xylR*	X57601	***GTTaGTTTAaTGgtTAAACAAAC***	−69	6.39	TTGAAAGCGaTTaatt	−110	4.01
Bacillus sp.	*xyn1*	AF015445	*GTTTGTTTAcTAgATAAACtAAg*	−177	6.34	aTGAAAGCGaaTTCAA	−140	4.53
						aTGAAAGCctTTaCAt	−107	4.44
B. halodurans	*xylAB/xylR*	C.G.	*GTTTGTcTATTGAATAAACtAAg*	−100	6.39	TTGAtAaCGCTTaCtt	−32	4.61
	BH2120 (xyn1)	C.G.	*GTTTGTTcAcTGgATcAACtAAg*	−190	5.68	aTGAAAGCcCTTTCtt	−101	4.30
	BH0700/BH0701-2-3-glcA-uxaC-uxuAB	C.G.	*cTTTGTTTgTTAAcTtAACAAAC*	−144	5.62	TTGtAAaCctTTTCAA	−77	4.45
	BH3678-79-80-81-82-xynB	C.G.	cTTTGTTTAaTGgtaAAACAAAg	−165	5.65	agGAAAaCGCTTaCAA	5	4.73
S. xylosus	*xylAB*	X57599	***GTTTGTTTATTAAATtAACcAAC***	−44	6.30	TaGtAAGCGtTTaCAA	−89	4.84
C. difficile	*xylBA/xylR*	U.G.	*GTTaGTTTAaTAtActAACAAAa*	−101	5.61	–		
C. difficile	*xylS-pts1-2-3-4*	U.G.	GTTaGTTTAaTGtAcAAAtgAAt	−229	4.98			
E. faecalis	*xylR/xylS-pts1-2-3-4-xylAB*	U.G.	*aTTTGTTTcaCgATAAACtAAt*	−233	5.60	aTacAAaCGCTTTCAt	−295	4.17
T. ethanolicus	*xylAB*	AF001974	***GTTTGTTgAaTCAATAAACtAtt***	−133	5.83	–		
	Consensus:		**GTTwGTTtatnnnataACwAAC**					
L. brevis	*xylAB*	AF045552	***GTTGGTTGTcgCAAgCAACtAAC***	−43	6.73	–		
	xylT	AF045552	***GTTGGTTGTgaCAAtCAACCAAC***	−108	6.48	**aaGAAAaCGgTTTCAA**	−156	4.70
L. pentosus	*xylAB*	M57384	***GTTGGTTGccgAAAcaAACtAAC***	−39	6.22	TaGAAAGCGtTTaCAA	−89	4.90
	xylPQ	U89276	***aTTaGTTGTaTTcAACAACCAAC***	−91	6.46	–		
	Consensus:		**GTTgGTTGnnnnnnnCAACcAAC**					
C. acetobutylicum	*xynCB*	C.G.	*ACTTTTTAAAGtgCTTTTtAAAAGT*	−61	7.00	–		
			gCTTTTTAAAAGTtaTTgAAAAAGT	−49	6.81			
	xylB	C.G.	*ACTTaTTAAAGGAtaTTTAAAAAGT*	−291	7.09	aTGAAAaaGtTTaCAA	−345	4.37
	xylR (CAC3673)	C.G.	*ACTTTaTAAAAcAaCTTTAtAAAGg*	−98	6.66	–		
	Consensus:		**ACTTTTTAAAnnnnnTTTAAAAAGT**			**TTGAAAgCGcTTTCAA**		

Position of the first base in the site is given in column pos. The known regulatory sites are underlined. Capital letters, nucleotides conforming to the consensus. The known-regulated genes are shown in bold. The sites included in the training set are shown in italics. The site scores that are below of the used thresholds (4.86 for AraR and 4.1 for CcpA) are underlined. C.G. and U.G. in column AC stand for complete or unfinished genome, respectively.

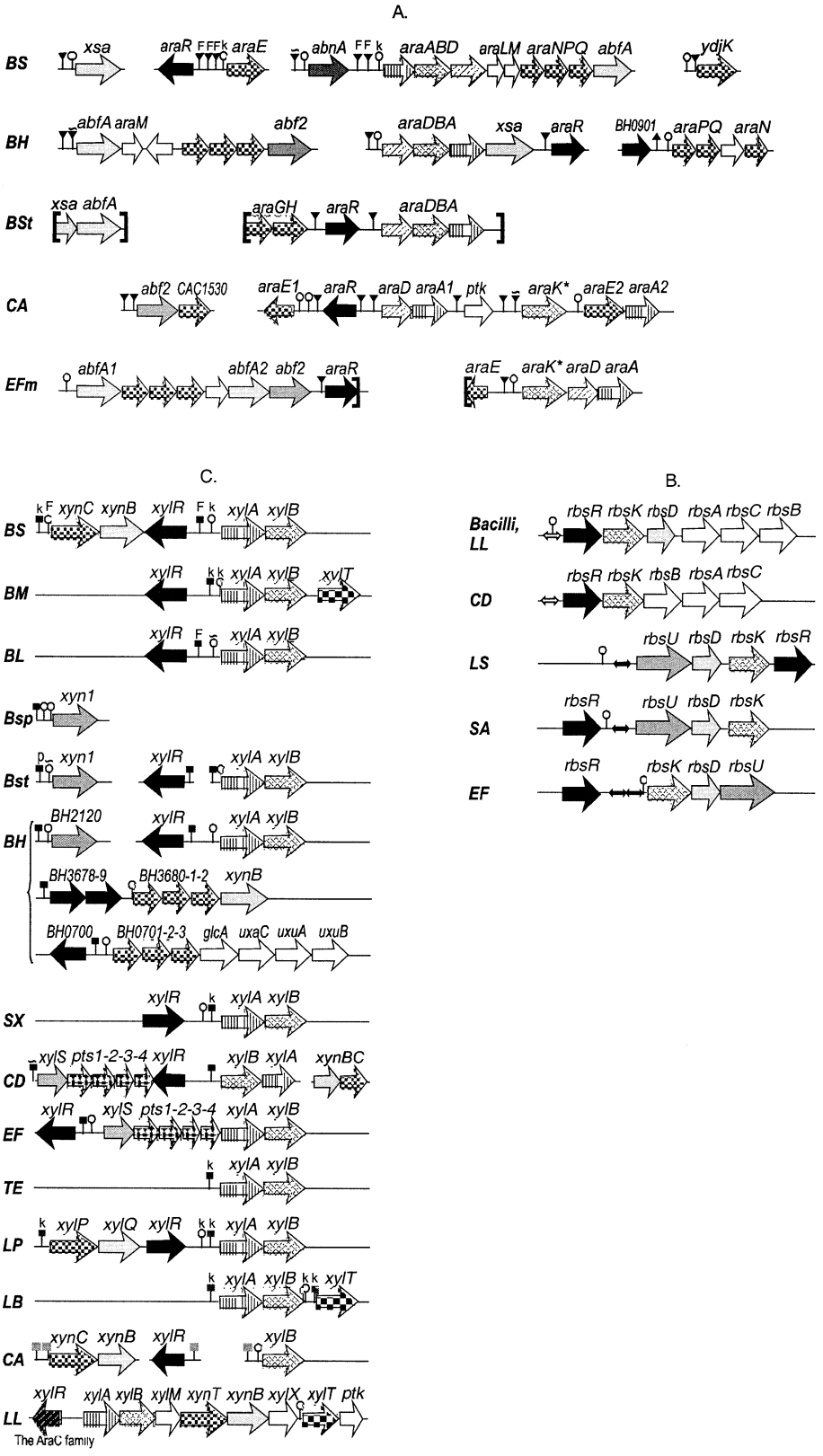

Fig. 1. The operon structures and the transcriptional regulatory sites of arabinose (A), ribose (B) and xylose utilisation genes (C) in *B. subtilis* (BS), *B. halodurans* (BH), *B. stearothermophilus* (st), *B. megaterium* (BM), *B. licheniformis* (BL), *Bacillus sp.* (B*sp*), *C. acetobutylicum* (CA), *C. difficile* (CD), *E. faecalis* (EF), *E. faecium* (E*fm*), *S. aureus* (SA), *S. xylosus* (SX), *T. ethanolicus*. (TE), *L. pentosus* (LP), *L. brevis* (LB), *L. sakei* (LS) and *L. lactis* (LL). Non-orthologous gene displacements of *araK* are shown by asterisks. The AraR, RbsR, XylR and CcpA sites are shown by triangles, two-directed arrows, squares and circle, respectively. The letters F, K and P with any of these site note the existence of a footprinting experiment proved this site (F), known regulation of downstream genes (K) or prediction of this site in a previous publications. The scores of sites with tilde are slightly below of used threshold. The contig ends are marked by square brackets.

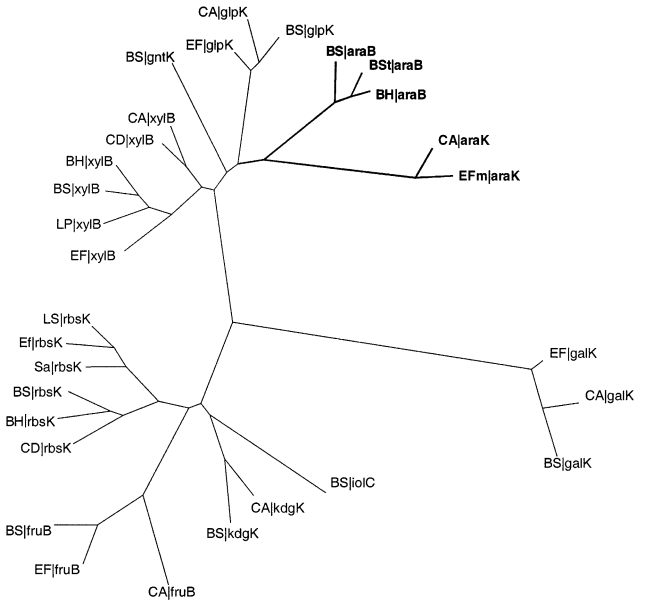

Fig. 2. A maximum likelihood phylogenetic tree of sugar kinases from the *Bacillus/Clostridium* group of bacteria. Bold, two distinct groups of arabinose kinases.

are likely to be co-regulated with the latter. In addition, we have identified one more putative member of the *C. acetobutylicum* AraR regulon, *abf2 (CAC1529),* which is highly similar to α-arabinofuranosidase 2 from *Streptomyces chartreusis* (52% identity). It is likely to form an operon with a hypothetical permease, *CAC1530*, which is homologous to the raffinose permease *rafB* and the lactose permease *lacY* from *Escherichia coli* (respectively, 44% and 41% identity). However, the *B. halodurans* genome contains a hypothetical gene *BH1867* orthologous to *abf2*. It is located in immediate proximity to the *araDBA-xsa* operon and the *abfA* gene, but does not have upstream candidate AraR sites.

The predicted AraR regulon in *E. faecium* includes the *araKDA* and *araR* operons. Two paralogues of *abfA* that are organised in one putative operon with *abf2*, and the hypothetical ABC transport system have no AraR site. Nevertheless, they are positionally clustered with *araR*.

3.2. The ribose regulon

There are two different types of the *rbs* operon in bacteria from the *Bacillus/Clostridium* group. The gene order *rbsRKDACB* is conserved in *B. subtilis*, *B. halodurans*, *B. stearothermophilus*, *B. anthracis* and *L. lactis*. The more distant genome of *C. difficile* has a putative *rbsRKBAC* operon without *rbsD* (small transmembrane component of the ribose transporter). The second type is present in *L. sakei*, where the *rbs* operon encoding RbsK, RbsD and a new ribose transporter named RbsU is regulated by RbsR [35]. The same genes form the ribose gene cluster in *E. faecalis* and *S. aureus* (Fig. 1B). On the phylogenetic tree of the LacI/GalR family of transcriptional regulators, the RbsR regulators from the analysed bacteria fall into two distinct groups exactly corresponding to the two different types of the *rbs* operon (Fig. 3).

The training set of the *rbs* upstream regions from bacilli, *L. lactis* and *C. difficile* was used to construct the first search profile. The derived putative RbsR consensus, TtTaTGTAAcCGgTTACAtAaA, is highly similar to the CRE consensus, TTGAAAgCGcTTTCAA, but is 6 bp longer. The search for potential CRE boxes in the upstream regions of *rbs* shows that the predicted RbsR and CRE boxes coincide. However, the candidate sites are closer to the RbsR box profile than to the CRE box profile. (Table 1B). This means that the negative regulation by the global regulator CcpA and by the specific repressor RbsR is encoded by the same binding sites. This make sense, as the ribose regulon should be repressed both in the absence of ribose and in the presence of other, preferred, carbon sources.

The second search profile was constructed using the *rbs* upstream regions from *L. sakei*, *E. faecalis* and *S. aureus*. The resulting RbsR consensus differs from the first one and from the CRE box consensus. However, in these genomes the candidate CRE boxes are located close to the candidate RbsR boxes in the *rbs* upstream regions.

Finally, the search with the derived RbsR profiles has not revealed any new candidate members of the RbsR regulons. Thus, the RbsR regulators seems to be true local repressors influencing the expression of only one operon in any genome.

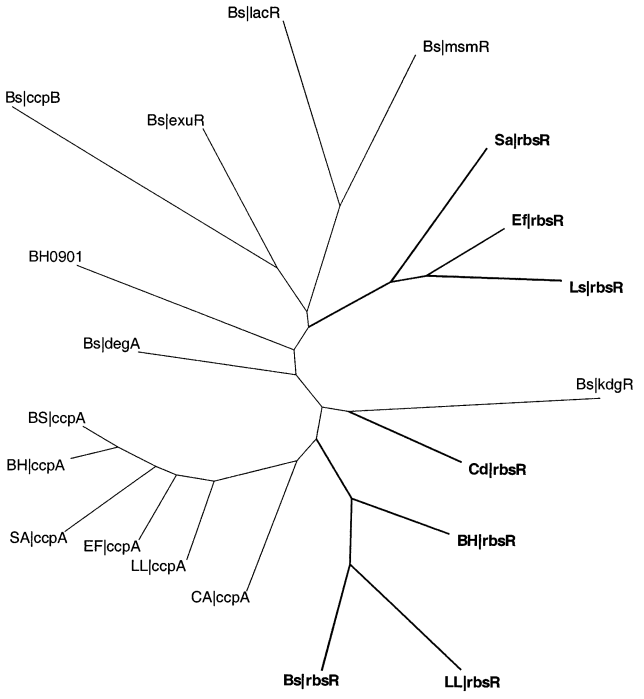

Fig. 3. A maximum likelihood phylogenetic tree of the LacI-family transcriptional regulators from the *Bacillus/Clostridium* group of bacteria. The tree includes all CcpA and RbsR proteins considered in this work, putative regulator BH0901 from *B. halodurans* and the others LacI-family regulators from *B. subtilis*. Bold, ribose repressors.

3.3. The xylose regulon

All operons orthologous to *xylAB*, *xynCB*, *xylR* and *xyn1* were collected from GenBank as well as from complete and unfinished genome sequences. The orthologues of the *xylAB* operon were found in four *Bacillus* and two *Clostridium* genomes and in the *E. faecalis*, *S. xylosus* [36], *T. ethanolicus* [37], *Lactobacillus pentosus* [38], *Lactobacillus brevis* [39] and *L. lactis* genomes (Fig. 1C). The orthologues of *xynB* were found in *B. halodurans*, *L. pentosus*, *L. lactis* and two *Clostridium* genomes. Two orthologues of *xyn1* from *B. stearothermophilus* were found in *Bacillus* sp. and *B. halodurans*.

The XylR regulators from Gram-positive bacteria belong to the so-called ROK family (repressor, ORF, kinase). The phylogenetic tree can be divided into two large groups and a separate branch corresponding to the XylR orthologue from *C. acetobutylicum* (Fig. 4). The first group consists of the xylose repressors from *Bacillus* genomes, whereas the second one includes XylR from the rest of the analysed Gram-positives, excluding *L. lactis.* The latter has a xylose regulator, also called XylR, from the AraC family of transcription factors.

Then, the signal determination procedure was applied to the training set consisting of upstream regions of the collected operons. Two highly similar signals were identified. The first signal is common for four xylose operons from *L. pentosus* and *L. brevis*, whereas the second one includes sites from the remaining genomes, except *L. lactis* and *C. acetobutylicum.* The corresponding 23-bp consensus sequences, GTTgGTTGnnnnnnnCAACcAAC and GTTwGTTtatnnnataAACwAAC, are similar to the XylR consensus earlier described for *B. subtilis* [17]. The search with the second constructed profile using the threshold of 5.6 (the minimal score of 14 sites from the training set) identified some new candidate members of the xylose regulon (Table 1C).

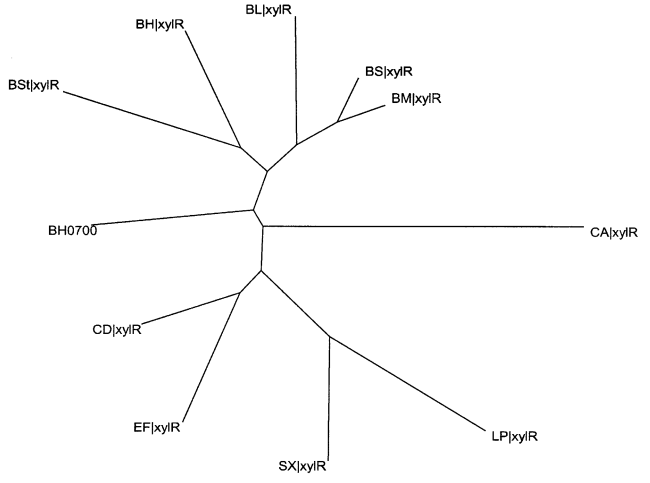

Fig. 4. A maximum likelihood phylogenetic tree of the transcriptional regulators from the ROK family including all known and predicted xylose repressors from the *Bacillus*/*Clostridium* group of bacteria.

A high-score XylR site occurs upstream of the hypothetical *BH3678-79-80-81-81-xynB* operon of *B. halodurans*. The genes of this operon encode a two-component regulatory system, an ABC system and the xylosidase XynB, respectively. The transport system is homologous to various disaccharide ABC transporters from the MalF-G-E-K subfamily. Based on the observation that *xynB* in all cases forms a single operon with transporters, we propose that this ABC system is an active transporter of xylosides. *BH0701*, a paralogue of *xylR* (36% identity), is transcribed divergently from the operon encoding a hypothetical transport system, the glucosidase GlcA and three enzymes of the glucuronate catabolism, UxaC, UxuA and UxuB. The common regulatory region of these operons contains a candidate XylR site that in fact can be the target of the local regulator BH0701.

The predicted XylR site in *E. faecalis* is located in the common regulatory region of the divergon formed by *xylR* and a putative operon including *xylAB*. The first five genes of these operons, encoding a putative xylosidase and a PTS transport system, were named *xylS* and *pts1-2-3-4*, respectively. The PTS transport system may be involved in the xylose- or xyloside-specific transport in *E. faecalis*. Note that in *C. difficile*, the orthologous *xylS-pts1-2-3-4* operon is located closely to the *xylAB*/*xylR* divergon and has a weak predicted XylR site (the score equals 4.98). There are no candidate XylR sites upstream of the *xynBC* operon.

In *C. acetobutylicum*, the upstream regions of *xynBC*, *xylB*, and *xylR* operons have no sequences resembling the XylR signal. Since the putative xylose repressor (CAC3673) from *C. acetobutylicum* is only distantly related to other XylR (Fig. 1C), we have tried to detect the original regulatory signal. Using the signal determination procedure we detected a common signal for the *xynCB, xylB* and *xylR* operons from *C. acetobutylicum* with the 23-bp consensus sequence ACTTTTTAAAnnnnnTTTAAAAAGT. Based on the fact that the above highly conserved signal occurs only upstream of the xylose metabolism genes and the putative XylR-related regulatory gene in *C. acetobutylicum* (Table 1C), we propose that XylR (CAC3673) is the regulator of the xylose regulon in *C. acetobutylicum*.

Interestingly, the obtained consensus of the XylR site is similar to the signal STTATTTnnnnnnnnnAAATAAS of the *N*-acetyl-glucose-amine repressor NagC from *E. coli*, also belonging to the ROK family [40].

3.4. The catabolite repression by CcpA

The CRE search profile was constructed using the set of known CRE sites from *B. subtilis*. The minimum site score from the training set was chosen as the threshold for the site search. Scanning of all studied genomes with the CRE profile reveals that the majority of genes for the utilisation of pentoses have a predicted CRE box, and thus are likely

to be regulated by CcpA. The rare exceptions are *xylAB* of *T. ethanolicus*, *L. brevis* and *C. difficile*, as well as *xynCB* of *C. acetobutylicum*, *xylPQ* of *L. pentosus* and *xyn1* of *B. stearothermophilus*.

3.5. Conclusions

We have identified a number of a new genes encoding transporters and enzymes involved in the utilisation of pentoses. Non-orthologous gene displacements of ribulose kinase were predicted in *C. acetobutylicum* and *E. faecium*. Simultaneous analysis of several genomes of the *Bacillus/Clostridium* group allowed us to derive the new consensus signals for the XylR and RbsR repressors, to predict candidate AraR and XylR sites and, consequently, new members of the arabinose and xylose regulons. In addition, the CRE sites of the global catabolite repressor CcpA were found upstream of the most genes from the AraR, RbsR and XylR regulons. It confirms the dual regulation of these genes by two opposite mechanisms, substrate induction and catabolite repression. The comparison of these results with the same analysis of Gram-negative bacteria is presented in the accompanying paper [41].

Acknowledgements

We are grateful to A. Rakhmaninova and o. Laikova for useful discussion. This study was partially supported by Grants from INTAS (99-1476) and HHMI (55000309).

References

[1] Rodionov, D.A., Mironov, A.A., Rakhmaninova, A.B. and Gelfand, M.S. (2000) Transcriptional regulation of transport and utilization systems for hexuronides, hexuronates and hexonates in gamma purple bacteria. Mol. Microbiol. 38, 673–683.

[2] Mironov, A.A., Koonin, E.V., Roytberg, M.A. and Gelfand, M.S. (1999) Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes. Nucleic Acids Res. 27, 2981–2989.

[3] Rodionov, D.A., Gelfand, M.S., Mironov, A.A. and Rakhmaninova, A.B. (2001) Comparative approach to analysis of regulation in complete genomes: multidrug resistance systems in gamma-proteobacteria. J. Mol. Microbiol. Biotechnol. 3, 319–324.

[4] Gelfand, M.S., Koonin, E.V. and Mironov, A.A. (2000) Prediction of transcription regulatory sites in Archaea by a comparative genomic approach. Nucleic Acids Res. 28, 695–705.

[5] Gelfand, M.S. (1999) Recognition of regulatory sites by genomic comparison. Res. Microbiol. 150, 755–771.

[6] Gelfand, M.S., Novichkov, P.S., Novichkova, E.S. and Mironov, A.A. (2000) Comparative analysis of regulatory patterns in bacterial genomes. Brief. Bioinform. 1, 357–371.

[7] McCue, L., Thompson, W., Carmack, C., Ryan, M.P., Liu, J.S., Derbyshire, V. and Lawrence, C.E. (2000) Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. Nucleic Acids Res. 29, 774–782.

[8] Robison, K., McGuire, A.M. and Church, G.M. (1998) A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K-12 genome. J. Mol. Biol. 284, 241–254.

[9] Stojanovic, N., Florea, L., Riemer, C., Gumuchio, D., Slightom, J., Goodman, M., Miller, W. and Harrison, R. (1999) Comparison of five methods for finding conserved sequences in multiple alignments of gene regulatory regions. Nucleic Acids Res. 27, 3899–3910.

[10] Florea, L., Riemer, C., Schwartz, S., Zhang, Z., Stojanovic, N., Miller, W. and McClelland, M. (2000) Web-based visualization tools for bacterial genome alignments. Nucleic Acids Res. 28, 3486–3496.

[11] Stulke, J. and Hillen, W. (2000) Regulation of carbon catabolism in *Bacillus* species. Annu. Rev. Microbiol. 54, 849–880.

[12] Mota, L.J., Tavares, P. and Sa-Nogueira, I. (1999) Mode of action of AraR, the key regulator of L-arabinose metabolism in *Bacillus subtilis*. Mol. Microbiol. 33, 476–489.

[13] Krispin, O. and Allmansberger, R. (1998) The *Bacillus subtilis* AraE protein displays a broad substrate specificity for several different sugars. J. Bacteriol. 180, 3250–3252.

[14] Sa-Nogueira, I., Nogueira, T.V., Soares, S. and de Lencastre, H. (1997) The *Bacillus subtilis* L-arabinose (*ara*) operon: nucleotide sequence, genetic organization and expression. Microbiology 143, 957–969.

[15] Woodson, K. and Devine, K.M. (1994) Analysis of a ribose transport operon from *Bacillus subtilis*. Microbiology 140, 1829–1838.

[16] Schmiedel, D., Kintrup, M., Kuster, E. and Hillen, W. (1997) Regulation of expression, genetic organization and substrate specificity of xylose uptake in *Bacillus megaterium*. Mol. Microbiol. 23, 1053–1062.

[17] Dahl, M.K., Degenkolb, J. and Hillen, W. (1994) Transcription of the *xyl* operon is controlled in *Bacillus subtilis* by tandem overlapping operators spaced by four base-pairs. J. Mol. Biol. 243, 413–424.

[18] Scheler, A. and Hillen, W. (1994) Regulation of xylose utilization in *Bacillus licheniformis*: Xyl repressor–*xyl*-operator interaction studied by DNA modification protection and interference. Mol. Microbiol. 13, 505–512.

[19] Lindner, C., Stulke, J. and Hecker, M. (1994) Regulation of xylanolytic enzymes in *Bacillus subtilis*. Microbiology 140, 753–757.

[20] Gat, O., Lapidot, A., Alchanati, I., Regueros, C. and Shoham, Y. (1994) Cloning and DNA sequence of the gene coding for *Bacillus stearothermophilus* T-6 xylanase. Appl. Environ. Microbiol. 60, 1889–1896.

[21] Kraus, A., Hueck, C., Gartner, D. and Hillen, W. (1994) Catabolite repression of the *Bacillus subtilis xyl* operon involves a *cis* element functional in the context of an unrelated sequence and glucose exerts additional XylR-dependent repression. J. Bacteriol. 176, 1738–1745.

[22] Galinier, A., Deutscher, J. and Martin-Verstraete, I. (1999) Phosphorylation of either Crh or HPr mediates binding of CcpA to the *Bacillus subtilis xyn cre* and catabolite repression of the *xyn* operon. J. Mol. Biol. 286, 307–314.

[23] Rygus, T. and Hillen, W. (1992) Catabolite repression of the *xyl* operon in *Bacillus megaterium*. J. Bacteriol. 174, 3049–3055.

[24] Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A. and Wheeler, D.L. (2000) GenBank Nucleic Acids Res. 28, 15–18.

[25] Schneider, T.D., Stormo, G.D., Gold, L. and Ehrenfeucht, A. (1986) Information content of binding sites on nucleotide sequences. J. Mol. Biol. 188, 415–431.

[26] Mironov, A.A., Vinokurova, N.P. and Gelfand, M.S. (2000) GenomeExplorer: software for analysis of complete bacterial genomes. Mol. Biol. 34, 222–231.

[27] Tatusov, R.L., Galperin, M.Y., Natale, D.A. and Koonin, E.V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 28, 33–36.

[28] Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

[29] Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. and Higgins, D.G. (1997) The CLUSTAL_X windows interface: flexible

strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25, 4876–4882.
[30] Felsenstein, J. (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. 17, 368–376.
[31] Page, R.D. (1996) TreeView: an application to display phylogenetic trees on personal computers. Comput. Appl. Biosci. 12, 357–358.
[32] Paulsen, I., Sliwinski, M. and Saier, M. (1998) Microbial genome analyses: global comparisons of transport capabilities based on phylogenies, bioenergetics and substrate specificities. J. Mol. Biol. 277, 573–592.
[33] Sa-Nogueira, I. and Ramos, S.S. (1997) Cloning, functional analysis, and transcriptional regulation of the *Bacillus subtilis araE* gene involved in L-arabinose utilization. J. Bacteriol. 179, 7705–7711.
[34] Shulami, S., Gat, O., Sonenshein, A. and Shoham, Y. (1999) The glucuronic acid utilization gene cluster from *Bacillus stearothermophilus* T-6. J. Bacteriol. 181, 3695–3704.
[35] Stentz, R. and Zagorec, M. (1999) Ribose utilization in *Lactobacillus sakei*: analysis of the regulation of the *rbs* operon and putative involvement of a new transporter. J. Mol. Microbiol. Biotechnol. 1, 165–173.
[36] Sizemore, C., Wieland, B., Gotz, F. and Hillen, W. (1992) Regulation of *Staphylococcus xylosus* xylose utilization genes at the molecular level. J. Bacteriol. 174, 3042–3048.
[37] Erbeznik, M., Dawson, K.A. and Strobel, H.J. (1998) Cloning and characterization of transcription of the *xylAB* operon in *Thermoanaerobacter ethanolicus*. J. Bacteriol. 180, 1103–1109.
[38] Chaillou, S., Lokman, B.C., Leer, R.J., Posthuma, C., Postma, P.W. and Pouwels, P.H. (1998) Cloning, sequence analysis and characterization of the genes involved in isoprimeverose metabolism in *Lactobacillus pentosus*. J. Bacteriol. 180, 2312–2320.
[39] Chaillou, S., Bor, Y.C., Batt, C.A., Postma, P.W. and Pouwels, P.H. (1998) Molecular cloning and functional expression in *Lactobacillus plantarum 80* of *xylT*, encoding the D-xylose-H^+ symporter of *Lactobacillus brevis*. Appl. Environ. Microbiol. 64, 4720–4728.
[40] Plumbridge, J. (2001) DNA binding sites for the Mlc and NagC proteins: regulation of *nagE*, encoding the *N*-acetylglucosamine-specific transporter in *Escherichia coli*. Nucleic Acids Res. 29, 506–514.
[41] Laikova, O.N., Mironov, A.A. and Gelfand, M.S. (2001) Transcriptional regulation of pentose utilization systems in the gamma subdivision of *Proteobacteria*. FEMS Microbiol. Lett. 205, 315–322.