ELSEVIER

FEMS Microbiology Letters 10239 (2001) 1–8

FEMS MICROBIOLOGY Letters

www.fems-microbiology.org

Computational analysis of the transcriptional regulation of pentose utilization systems in the gamma subdivision of *Proteobacteria*

Olga N. Laikova [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

Abstract

The comparative approach to the recognition of transcription regulatory sites is based on the assumption that as long as a regulator is conserved in several genomes, one can expect that sets of co-regulated genes (regulons) and regulatory sites for the regulator in these genomes are conserved as well. We used this approach to analyze the ribose (RbsR), arabinose (AraC), and xylose (XylR) regulons of gamma *Proteobacteria* for which (almost) completely sequenced genomes were available. Candidate binding sites for RbsR and AraC were detected. The improved XylR site consensus was proposed. Potential new members of the xylose regulons were found in the *Escherichia coli*, *Salmonella typhi*, and *Klebsiella pneumoniae* genomes. The function of these new xylose-regulated operons is likely to be the utilization of oligosaccharides containing xylose. Finally, candidate cAMP receptor-protein sites were identified in the regulatory regions of the majority of RbsR-, AraC-, and XylR-regulated operons. © 2001 Published by Elsevier Science B.V. on behalf of the Federation of European Microbiological Societies.

Keywords: Computer analysis; Transcriptional regulation; RbsR; AraC; XylR; cAMP receptor protein

1. Introduction

Each of the commonest pentoses, D-ribose, D-xylose and L-arabinose, can support the growth of *Escherichia coli* on a mineral medium as the sole source of carbon and energy. The systems for utilization of these sugars are regulated on the transcriptional level by protein factors which belong to different structural families: the ribose regulator (RbsR) belongs to the LacI family, whereas the arabinose (AraC) and xylose (XylR) regulators are members of the AraC family. Members of these families have the opposite modes of action: RbsR is a classical repressor whose true inducer is D-ribose, whereas AraC and XylR are primarily activators of transcription. In addition to the control by the local transcriptional regulators, the pentose transport and feeder metabolic pathways are modulated by the global transcription regulator CRP (cAMP receptor protein) [1–4].

Uptake of ribose, xylose, or arabinose can be mediated by high-affinity transporters dependent on periplasmic binding proteins specific for the respective sugars, that is, the ABC (ATP-binding cassette) -type transporters. In addition, *E. coli* has low-affinity transporters for xylose (XylE) and arabinose (AraE) powered by the proton gradient. It is remarkable that even the high-affinity xylose permease is able to transport ribose and can substitute for the eliminated ribose transport system [5].

Internally, the pentoses are converted to the intermediate compounds of the non-oxidative part of the pentose phosphate pathway. One step, catalyzed by ribokinase RbsK (EC 2.7.1.15), leads from ribose to D-ribose-5-phosphate, whereas both D-xylose and L-arabinose are turned into D-xylulose-5-phosphate (in two or three steps respectively). Intracellular xylose is converted by xylose isomerase XylA (EC 5.3.1.5) to D-xylulose, which in turn is phosphorylated by xylulokinase XylB (EC 2.7.1.17). The three consecutive steps of the L-arabinose metabolism are catalyzed by L-arabinose isomerase AraA (EC 5.3.1.4), L-ribulokinase AraB (EC 2.7.1.16) and L-ribulose-phosphate epimerase AraD (EC 5.1.3.4).

The genes encoding the ribose transporter and ribokinase form an inducible operon *rbsDACBK*. The regulator gene for this operon (*rbsR*) immediately follows the *rbsK* gene, is transcribed in the same direction, and seems to

* Corresponding author. Fax: +7 (95) 315-0501.
E-mail address: laikova@mail.ru (O.N. Laikova).

0378-1097/01/$20.00 © 2001 Published by Elsevier Science B.V. on behalf of the Federation of European Microbiological Societies.
PII: S0378-1097(01)00495-5

represent a separate transcriptional unit. In *E. coli*, RbsR was experimentally shown to bind to a palindromic sequence upstream of the *rbs* operon [6].

The *E. coli* growth on arabinose involves expression of three unlinked L-arabinose-inducible operons, one of which encodes enzymes (*araBAD*) and the other two encode systems of the arabinose uptake (*araFGH* and *araE*). The regulator gene *araC* is transcribed divergently from the *araBAD* operon. In the presence of L-arabinose, AraC acts as a transcriptional activator for the *araBAD*, *araFGH* and *araE* promoters, as well as for the *araJ* promoter serving a gene of an obscure function. On the other hand, in the absence of arabinose, AraC represses the *araBAD* promoter. It also represses the promoter of its own gene both in the absence and presence of arabinose. The mechanisms of the transcription regulation by AraC have been intensely studied for over 40 years (see ref. [7] for a review). According to the currently accepted model, the AraC protein functions as a dimer, with each subunit recognizing a 17-bp half-site. In natural operators, these half-sites occur as direct repeats [8,9]. When arabinose is absent, the AraC dimer prefers binding to half-sites separated by a long stretch of DNA ($\sim$200 bp), thus forming a DNA loop. On addition of arabinose, AraC tends to bind adjacent half-sites with a spacer of 4 bp [10]. The AraC-binding sites in *E. coli* and *Salmonella typhimurium* were characterized by footprinting and mutational analyses [1–3,8,9,11–13].

Two inducible loci are responsible for the xylose utilization in *E. coli*. One of these loci is represented by the single-gene operon *xylE*, and the other one comprises the *xylAB* and *xylFGH* operons, the latter encoding the high-affinity transporter. The intergenic region contains two divergently oriented promoters. The xylose regulator is encoded by the *xylR* gene situated downstream of the *xylFGH* operon. This gene is transcribed in the same direction and has its own weak promoter. It was experimentally demonstrated that XylR acts as a transcriptional activator for the promoters of the *xylAB* and *xylFGH* operons, but not for the *xylR* promoter. Two regions of enhanced XylR binding in the presence of xylose were revealed in the *xylAB*/*xylFGH* intergenic space by DMS footprinting [4].

The natural source of xylose is the plant cell wall material, where it is the principal component of hemicelluloses, mostly xylans and xyloglucans. These heteropolymers consist of a backbone formed by α-(1,4)-linked residues of D-xylose (in xylans) or D-glucose (in xyloglucans) and various branching saccharidic groups. The xylan side chains may contain such sugars as arabinose or glucuronic acid, and the backbone of xyloglucans mainly carries α-(1,6)-linked D-xylose residues. The structural diversity of hemicelluloses implies a wide range of enzymatic activities required to completely destroy the polymers. Given the complexity of the degradation of the plant material, multispecies communities are usually involved in this process. The members of these communities greatly differ in their capacities to hydrolyze the initial polymers or the products of their partial breakdown [14]. As examples of such intermediates, one can mention α-(1,4)-xylooligosaccharides, down to xylobiose, and isoprimeverose (α-D-xylopyranosyl-(1,6)-D-glucopyranose), a product of xyloglucans hydrolysis.

Using the comparative approach (see the accompanying paper [15] for the details), we characterized the ribose, arabinose, and xylose regulons in genomes of several gamma *Proteobacteria*. We revised and improved the XylR site consensus and predicted some new potential members of the xylose regulons in the genomes of enteric bacteria. These genes are likely to be responsible for the utilization of xylooligosaccharides.

2. Materials and methods

The following complete genome sequences were downloaded from GenBank database [16]: *E. coli* K-12 (accession number U00096), *Haemophilus influenzae* (L42023), *Pasteurella multocida* (AE004439), *Pseudomonas aeruginosa* (AE004091), and *Vibrio cholerae* (AE003852 and AE003853). The complete sequences of the *Salmonella typhi* and *Yersinia pestis* genomes were obtained from the Sanger Centre web server (http://www.sanger.ac.uk), preliminary sequence data for *Klebsiella pneumoniae* and *Actinobacillus actinomycetemcomitans* were downloaded from the web sites of the Genome Sequencing Center at the Washington University (http://genome.wustl.edu) and the Advanced Center for Genome Technology of the University of Oklahoma (http://www.genome.ou.edu) respectively.

The protein and nucleotide databases were searched at the NCBI web site (http://www.ncbi.nlm.nih.gov) using BLAST [17]. The homology relationships of proteins were determined with the help of the InterPro database [18]. In addition, the glycoside hydrolases classification (http://www.expasy.ch/cgi-bin/lists?glycosid.txt) and the transport protein classification (http://www-biology.ucsd.edu/–msaier/transport) presented on the web were used. The putative genes in the unpublished genomes can be accessed using open reading frame identifiers assigned in the ERGO database (http://wit.integratedgenomics.com/igwit). Genomic analyses were done using GenomeExplorer [19]. Binding signals of transcription factors were determined using SignalX [20]. The recognition profile for CRP based on the sites collected from the literature was kindly provided by D.A. Rodionov. For other details see the accompanying paper [15].

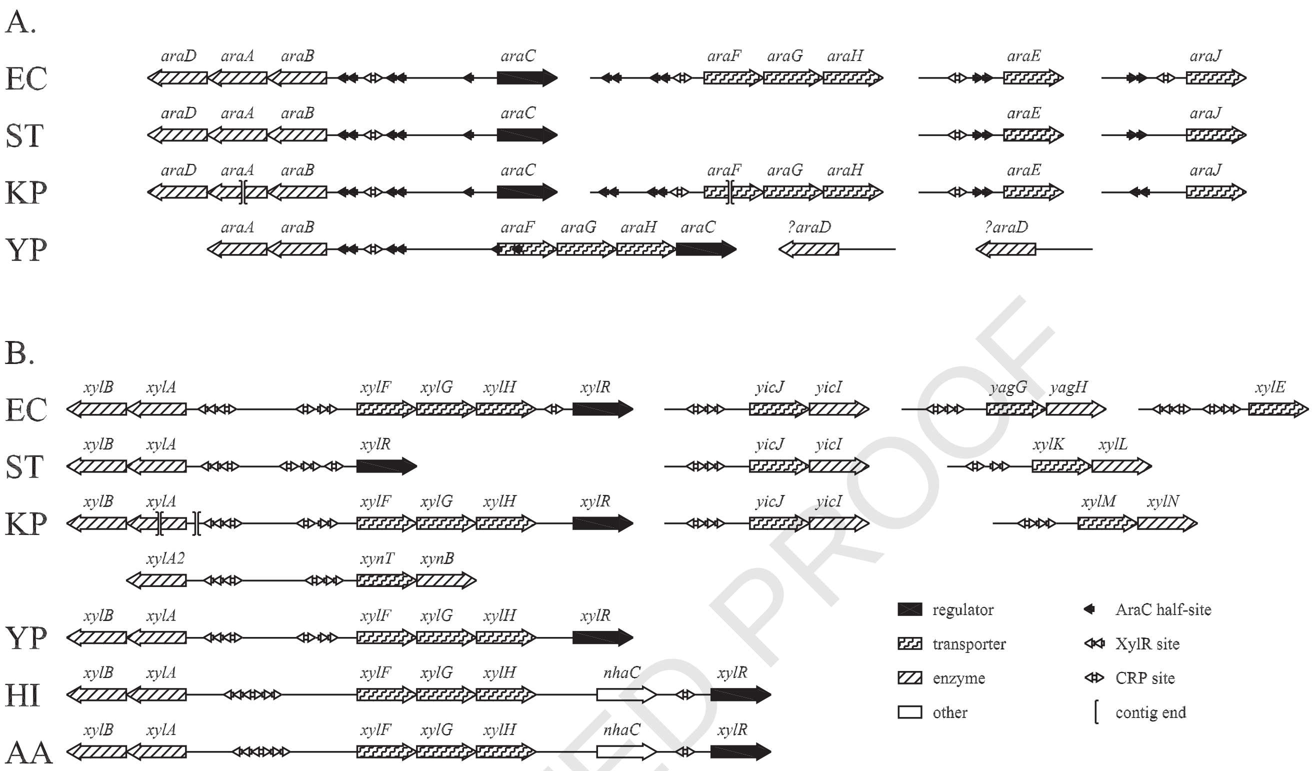
A.
EC
araD araA araB araC araF araG araH araE araJ
ST
araD araA araB araC araE araJ
KP
araD araA araB araC araF araG araH araE araJ
YP
araA araB araF araG araH araC ?araD ?araD
B.
EC
xylB xylA xylF xylG xylH xylR yicJ yicI yagG yagH xylE
ST
xylB xylA xylR yicJ yicI xylK xylL
KP
xylB xylA xylF xylG xylH xylR yicJ yicI xylM xylN
xylA2 xynT xynB
YP
xylB xylA xylF xylG xylH xylR
HI
xylB xylA xylF xylG xylH nhaC xylR
AA
xylB xylA xylF xylG xylH nhaC xylR
regulator
transporter
enzyme
other
AraC half-site
XylR site
CRP site
contig end

Fig. 1. The operon structures and regulatory sites in the arabinose (A) and xylose (B) regulons of *E. coli* (EC), *S. typhi* (ST), *K. pneumoniae* (KP), *Y. pestis* (YP), *H. influenzae* (HI), and *A. actinomycetemcomitans* (AA). The common gene names correspond to the gene names in the *H. influenzae* genome as follows: *xylA* (*HI1112*), *xylB* (*HI1113*), *xylF* (*HI1111*), *xylG* (*HI1110*), *xylH* (*HI1109*), *xylR* (*HI1106*), *nhaC* (*HI1107*).

3. Results and discussion

3.1. RbsR regulon

Orthologs of the *E. coli rbsR* gene were found in the genomes of *S. typhi*, *K. pneumoniae*, *V. cholerae*, *H. influenzae* and *P. multocida*, along with the structural *rbs* genes arranged in the same order as in the *E. coli* genome (Fig. 1). In the *P. aeruginosa* genome, the *rbsD* gene is absent and the operon structure is clearly different, the gene order being *rbsBACRK*. The SignalX program was applied to the set of the ribose operon upstream regions and the candidate sites were detected in all species, the one in *E. coli* being previously known. However, the site found in *P. aeruginosa* showed clear systematic differences compared to all other candidate sites and conformed to the earlier-proposed consensus for the PurR-binding sites [21,22]. By that reason, the site in *P. aeruginosa* was not included in the training set for the RbsR-binding-site recognition profile. The genomes of *E. coli*, *S. typhi*, *K. pneumoniae*, *V.*

Table 1
The regulatory sites of the ribose regulons in gamma *Proteobacteria*

Genome	Operon	Site	Pos.	Score	CRP site	Pos.	Score
The RbsR regulon							
E. coli	CG *rbsDACBKR*	**TCAGCGAAACGTTTCGCTGA**	−31	6.96	cgtTtcGAggTtGATCACATTT	−102	3.86
S. typhi	CG *rbsDACBKR*	ctAGCGAAACGTTTCGAcGg	−111	6.08	cgtTtcGAcggcGATCACAaTT	−102	3.72
		cCAGCGAAACGTTTCGCTaq	−32	6.61			
K. pneumoniae	UG*rbsDACBKR*	ctcGCGAAACGTTTCGATGg	−116	6.08	cgtTtcGATggcGATCACATTT	−107	3.97
		TtAGCGAAACGTTTCGCTag	−38	6.71			
Y. pestis	CG *rbsDK*	TtAGCGAAACGTTTCGCTct	−31	6.17	tgtTtcGgTggcGATCACAaTT	−100	3.68
H. influenzae	CG *HI0501-HI0506*	TtATCGAAACGTTTCGATaA	−37	6.86	tttTGTGATCaAtATCcCAaaT	−105	4.75
P. multocida	CG *rbsDA_1C_1B_1KR*	TCATCGAAACGTTTCGATGA	−46	6.96	ttATtTGATCcAGtTCACAgaT	−120	4.76
V. cholerae	CG *VCA0127-VC0132*	TCATCGAAACGTTTCGATGt	−86	6.67	tttgcTGATCgtttTCACAcTc	−151	3.40
P. aeruginosa	CG *rbsBACRK*	taACGCAAACGTTTGCGTct	−76				

The known sites are shown in bold. The sites included in the training set are underlined. The bases which conform to consensus sequences are indicated by capital letters, except the site in *P. aeruginosa*, where capitals indicate bases forming palindrome. The CG and UG abbreviations in the second column indicate complete and unfinished genomes respectively.

cholerae, *H. influenzae* and *P. multocida* were scanned using the constructed profile, but no potential new members of the RbsR regulon were found. A curious situation was observed in the complete genome of *Y. pestis*, where the *rbsR* gene was definitely lost. All that is preserved of the *rbs* locus is the *rbsD* ortholog (68% identity) followed by *rbsK* (72% identity). The flanking regions of the locus are the same as in the *E. coli* genome. The presence of a high-score RbsR candidate site upstream of *rbsD* (Table 1) can be explained by the conjecture that the *rbsR* regulator gene has been deleted quite recently, so that the sequence drift has not yet destroyed the site.

3.2. *AraC regulon*

Orthologs of the AraC regulator were found only in *Enterobacteriaceae*, so the complete genomes of *E. coli*, *S. typhi*, *Y. pestis* and the unfinished genome of *K. pneumoniae* were considered. All known *E. coli* AraC-binding half-sites were collected and comprised the learning set for a two-box recognition profile. The profile was applied to the *S. typhi* and *Y. pestis* genomes. The high-score sites found in the upstream regions of the *ara* gene orthologs were added to the training set. The final AraC profile was constructed in two variants, the single-box and two-box profiles. The use of the single-box profile resulted in high overprediction, so it was used only to find single-box sites analogous to the O2 site at the *E. coli araBAD* promoter [11].

The AraC regulon is completely conserved in *K. pneumoniae*, whereas in *S. typhi*, the high-affinity transporter (*araFGH*) is lost. The peculiar feature is the inverse orientation of the candidate site upstream of *araJ* in *K. pneumoniae*. In the *Y. pestis* genome, the *araE* and *araJ* genes are missing and all other arabinose regulon genes, except *araD*, form one locus with the *araB*/*araF* intergenic region containing two potential AraC-binding sites. The strongest single half-sites which could correspond to the O2 operator are completely or partially overlapped by the *araF* coding sequence, and the distances between these half-sites and the relevant two-box site are substantially different from that in other genomes. It should be noted that the *araD* gene has a close homolog in the *E. coli* genome, *sgbE* (76% identity). Similarly, *Y. pestis* has two genes, unlinked to other *ara* genes, each having almost equal similarity both to *araD* and *sgbE*. Thus, it does not seem possible to unambiguously resolve the orthology relationships between the *E. coli* and *Y. pestis* genes. Neither of the *Y. pestis* genes has a significant candidate AraC site.

3.3. *XylR regulon*

The signal recognized by XylR was proposed previously [4], but the profile based on the suggested sites did not perform well. So, SignalX was applied to the footprinted regions. A signal whose structure is that of a 17-bp direct repeat with a 4-bp spacer was determined, which was in a better agreement with the experimental data (Fig. 2). Indeed, in the new version, the guanines protected by XylR against methylation [4] align well against each other. The profile was constructed using the newly determined sites and applied to the genomes of *E. coli*, *S. typhi*, *Y. pestis* and *H. influenzae*. Sites with the highest scores were found in the upstream regions of the operons orthologous to the *E. coli* genes *xylAB* and *xylFGH*, as well as upstream of the *E. coli* gene *xylE* which has no orthologs in the other genomes. All these sites were added to the learning set used to construct the final profile. This profile was used to scan the above-mentioned genomes and also the unfinished genomes of *K. pneumoniae* and *A. actinomycetemcomitans*. The absence of the autoregulation of *xylR* in *E. coli* was shown experimentally [4]. Our analysis showed the absence of candidate XylR-binding sites upstream of *xylR* in all genomes except *S. typhi*. In this genome, the *xylFGH* operon is lost, so the *xylA*/*xylF* intergenic region of *E. coli* corresponds to the *xylA*/*xylR* intergenic region of *S. typhi*. The latter region contains two candidate XylR-binding sites. The question of whether the site preceding the *xylR* gene has any functional significance can be answered only experimentally.

Search with the XylR profile and comparison of genes having strong candidate sites in the upstream regions have lead to the identification, in the *E. coli*, *S. typhi* and *K. pneumoniae* genomes, of putative operons, which are likely to be regulated by XylR and whose products probably mediate transport and hydrolysis of xylooligosaccharides. All these operons have some common features. They consist of two genes, the proximal one encoding a transporter, and the distal one encoding a putative glycoside hydrolase (EC 3.2.1.x). The transporters are homologous to each

```
                              *         -35
Ec_xylA   5'-gcgcacacttGTGAATTATCTCAATAGcagtGTGAAATAACATAATTGagcaactgaa-3'
Ec_xylA   3'-cgcgtgtgaaCACTTAATAGAGTTATCgtcaCACTTTATTGTATTAACtcgttgactt-5'

                    *                   -35
Ec_xylF   5'-caagaaataaACCAAAAATCGTAATCGaaagATAAAAATCTGTAATTGtttcccctg-3'
Ec_xylF   3'-gttctttattTGGTTTTTAGCATTAGCtttcTATTTTTAGACATTAACtttggggac-5'
```

Fig. 2. The XylR-binding sequences in the *xylA*/*xylF* intergenic region of *E. coli* (both DNA strands are shown). The XylR sites proposed previously [4] are underlined and the corrected sites are shown in capital letters. The predicted −35 elements of the xylA and xylF promoters are in bold and signed. The experimental data [4] are indicated: the guanines protected from the DMS methylation by XylR are shaded and the adenines whose methylation was enhanced in the presence of XylR are shaded and marked with asterisks.

other and are similar (25–36% identity) to the α-xyloside transporter XynC (*ynaJ*) of *Bacillus subtilis* and the isoprimeverose transporter XylP of *Lactobacillus pentosus* [23]. So, these transporters apparently belong to the galactosides–pentosides–hexuronides family, also known as the sodium:galactoside symporter family, of transporters (InterPro accession number IPR001927) [24].

The enzymes are more diverse and can be divided into two groups which represent different families of glycoside hydrolases, as classified by sequence similarity [25]. YagH of *E. coli*, and XynB and XylN of *K. pneumoniae* fall into the family 43. YagH of *E. coli* and XynB of *K. pneumoniae* are homologous (52 and 44% identity respectively) to the xylan 1,4-α-xylosidase (EC 3.2.1.37) XynB of *Bacillus pumilus* (Swiss-Prot accession number P07129). On the other hand, XylN of *K. pneumoniae* has no strong homologs, the closest one being the putative xylosidase of a ruminal anaerobe *Prevotella ruminicola* (34% identity) (DDBJ accession number BAA78558). This suggests a different catalytic activity. The hypothetical enzymes encoded by *yicI* of *E. coli*, with its orthologs in the *S. typhi* and *K. pneumoniae* genomes, and *xylL* of *S. typhi* represent family 31 of glycoside hydrolases (InterPro accession number IPR000322). This diverse family consists of proteins from Archaea, Eukaryota, and Bacteria that hydrolyze α-glycosidic bonds, including those formed by xylose residues. The closest experimentally characterized homolog of YicI is XylQ of *L. pentosus* (44% identity), an α-xylosidase highly specific for isoprimeverose. Note that the *xylPQ* operon of *L. pentosus* belongs to the xylose regulon of this Gram-positive bacterium [23]. Finally, although XylL of *S. typhi* has no close homologs with experimentally characterized function, the most similar protein encoded by the *xylS* gene of the archaeon *Sulfolobus solfataricus* (34% identity) exhibits the α-xylosidase activity [26].

The locus in the *K. pneumoniae* genome containing the previously mentioned *xynB* gene is specifically interesting. *K. pneumoniae* carries two copies of the *xylA* gene, one of which is transcribed divergently from the putative operon *xynTB*. Two quite strong candidate XylR-binding sites were found in the common intergenic promoter region. The most intriguing feature is the fact that XynT and XynB of *K. pneumoniae* are respectively 80 and 81% identical to the XynT and XynB of *Lactococcus lactis*. The genes encoding these products, in the same *xynTB* order, belong to the locus responsible for the xylose utilization in *L. lactis* [27]. This makes it likely that this operon has been subject to relatively recent horizontal transfer. However, the source and direction of this transfer will be unclear until more genomes harboring this operon are identified.

3.4. CRP modulation

Candidate CRP sites were found upstream of the operons of the RbsR, AraC and XylR regulons. The CRP profile is rather weak (has low specificity), so, in the cases when in the upstream region of an operon more than one potential CRP sites were found, we considered the site scores and the relative positions of the sites for CRP and the respective local regulator. In the xylose regulon, two variants of the relative location of the CRP and XylR sites were observed most frequently. In both cases, the CRP site is upstream of the XylR site, either immediately adjacent or separated by 10 bp. The genomes of *H. influenzae* and *A. actinomycetemcomitans* demonstrate an extreme form of the first variant, where the single CRP site is flanked with two XylR sites. The upstream region of the *xylKL* operon of *S. typhi* contains the CRP and XylR sites with an unusually long spacer of 21 bp. However, in most cases, the positions of CRP and XylR sites conform to the period of the DNA helix. This suggests some sort of functional interaction between CRP and XylR, e.g. co-operative binding.

4. Conclusions

The computational predictions done for *Proteobacteria* are generally based on the experimental data obtained for *E. coli*. So, the comparative approach to the prediction of transcription regulatory sites requires conservation of a regulator (experimentally studied in *E. coli*) in several sequenced genomes. When a well-studied regulator and the respective regulated genes are conserved only in closely related organisms, such as enterics, or a regulon is very simple, with only one regulated operon, the prediction amounts to identification of candidate binding sites for these operons. The arabinose and ribose regulons in *Proteobacteria* are such cases, and they have been analyzed in order to validate the approach and to provide the complete picture of the regulation of the pentose-utilization systems (Table 2). However, the continuing sequencing of bacterial genomes will certainly complicate even these relatively simple cases. Thus, the ribose regulons appear to be present in several *Pseudomonas* species and also in the *Burkholderia* group of beta *Proteobacteria*, and the RbsR-binding sites in those genomes are similar to the site in the *P. aeruginosa* genome, but not in the genomes analyzed here (data not shown). The study of the parallel evolution of the LacI family regulators and their binding sites is currently under way.

Still, the ribose regulons in the studied genomes are simple and uniform. Both in the *Bacillus*/*Clostridium* group and in gamma *Proteobacteria* they form one locus containing one regulated operon and the regulator gene from the LacI family. The structural genes are also similar, as they encode an ABC-type transporter and ribokinase. The only exception is *Y. pestis*. In this case, the ribokinase and the upstream region of the structural operon is retained, so the kinase expressed constitutively could allow the utilization of ribose taken up through another trans-

Table 2
The regulatory sites of the arabinose and xylose regulons in gamma *Proteobacteria*

Genome	Operon		Site	Pos.	Score	Dir.	CRP site	Pos.	Score
The AraC regulon									
E. coli	CG	*araBAD/araC*	**gAaaccAATtgTCCATA**	−310	3.17	⇒			
			TaaCAaAAgtgTCtATA-(4)-cgGCAGAAaAgTCCAcA	−173	7.90	⇒			
			TAGCAttTTtATCCATA-(4)-TAGCgGATcctaCCtgA	−99	8.13	⇒	ttATtTGcaCggcgTCACAcTT	−131	4.07
		araFGH	**cAaGGATTtccagGCTA-(4)-TATGGATTAATCTGCTg**	−284	8.31	⇐			
			TATGtcTTtTcCcGCTA-(4)-TATGcAcgtTctcaCTg	−209	7.19	⇐	cgATGTGATaTtGcTCtCcTaT	−163	3.67
		araE	**cAGCAatTTAATCCATA-(4)-TgctgttTccgaCCtgA**	−99	7.00	⇒	AatTGgaATaTccATCACATaa	−131	4.25
		araJ	**cAGCAGgATAATgaATA-(4)-ggGgcGAATtATCtcTt**	−120	6.26	⇒	tAtccTGcaagctATCACtTTa	−67	3.21
S. typhi	CG	*araBAD/araC*	gAaacaAATtgTCCATA	−311	3.48	⇒			
			TAaCAGAAgtgTCtATA-(4)-TgGCtGgAatgTCCAcA	−173	7.66	⇒			
			TAGCAttTTtgTCCATA-(4)-TAGCgGATcctgCCtgA	−99	8.11	⇒	AtATtTGcaCagcgTCACAcTT	−131	4.05
		araE	TAGCcatTTAATCCATA-(4)-TgcCgttTccAgCCtgA	−99	7.44	⇒	AAgTGTaATaTccATCACATaT	−131	4.78
		araJ	TAGCAtAATAATgaATA-(4)-aAGacGAATAAggaATg	−134	7.00	⇒			
K. pneumoniae	UG	*araBAD/araC*	ggcaccAATtgTCCATA	−309	2.82	⇒			
			TAaCAaAAgtgTCtATA-(4)-cgGCAGAAaAgTCCAcA	−172	7.90	⇒			
			cAGCAaAATAATCCATA-(4)-TAGCgaATccggCCtgA	−98	8.35	⇒	AtATtTGcaCggcgTCACAcTT	−130	4.08
		araFGH	cAaGGATaAgcCTGCTg-(4)-TATaGATgATcCTGCTA	−282	7.75	⇐			
			TATGtcaTtTTtTGCTA-(4)-TATGtAcgcAatTaCTg	−207	6.81	⇐	cgATGTGATaatGcTCtCgTaT	−161	3.63
		araE	cAGCAatATAgTCCATA-(4)-TgGCttATcctgCCtgA	−99	7.50	⇒	AagTtTaATgTccATCACAaaa	−131	4.41
		araJ	cATGaAcgAATCaGgct-(4)-TATGGAggAAaCTGCcg	−114	6.22	⇐			
Y. pestis	CG	*araBA/araFGHC*	TAGCtGAATgtgaCATA	−471	3.14	⇒			
			TAGttaATTtATgCATA	−423	3.97	⇒			
			cAGtttATTtATCCATA-(4)-cAGCAaATTtAaCCtcg	−189	7.72	⇒			
			aAGCAGAAaAgTgCATA-(4)-cAGCtatTggcTCCtcA	−114	6.97	⇒	ActTtTGActgAcATCACAaaa	−146	4.17
The AraC half-site consensus			**yAGCakaWtwrTCCATA**						
The XylR regulon									
E. coli	CG	*xylAB/xylFGHR*	GTGAAtTATCtcAATaG-(4)-GTGAAATAACaTAATTG	−112	10.21	⇒	tttTGcGAgCgAGcgCACAcTT	−132	4.13
		xylFGHR/xylAB	accAAAaATCGTAATcG-(4)-aTaAAAatctGTAATTG	−131	9.51	⇒	AAgTaaGATCTcGgTCAtAaaT	−163	3.73
		xylR					gttctTGATtTtGATaAaAaTT	−84	3.51
		xylE	tttTTACGTTATTTgtt-(4)-CAcTTACGATAaTTCtC	−219	8.52	⇐	tttcGTGcTCTgagTCACggca	−181	3.59
			aaGAcATtACGTAAacG-(4)-GTaAAAaATgaTAATTG	−107	8.94	⇒	AttTtgGATaattATCACAaTT	−141	4.17
							tAtcacaATtaAGATCACAgaa	−129	3.41
		yicJI	aacAAgaATCaTAAaTt-(4)-accAgATATCGgAATat	−109	7.04	⇒	tAAcGctAcCacGATCACATaa	−131	4.05
		yagGH	GcaAAATAACGTAATTc-(4)-aTaAgATATgaccATTG	−116	8.21	⇒	AtATaTcgTtggcgTCACAaaa	−137	2.97
S. typhi	CG	*xylAB/xylR*	GgGgAtTcTCtTAATaG-(4)-GTGAAATAACGTAATTG	−111	9.55	⇒	tttTGaGAgCcAGAgCACATTT	−133	4.49
		xylR/xylAB	acaAAAaAcCGTAATat-(4)-aTaAgAaATgacAATTG	−174	8.36	⇒	AAAaGcGATCgcGATCgaATcc	−206	3.06
							tgcTGTGATgaAttTCgCATaa	−124	4.16
		yicJI	aacAAgaATCGTAATTt-(4)-accAAATAACGgAATat	−101	8.00	⇒	tggcGctAcCctGATCACAgaa	−123	3.28
		xylKL	actAAtaAACGcAATcc-(4)-GTaAgATtTaaTAATTG	−196	7.63	⇒	AAcTGTGAcCacctTagCAaaT	−239	3.79

The divergently arranged operons are separated by a slash. The known sites are shown in bold. The sites included in the training sets are underlined. The bases which conform to consensus sequences are indicated by capital letters. The CG and UG abbreviations in the second column indicate complete and unfinished genomes respectively. The newly given gene names are *xylK* (ERGO identifier RTY04320) and *xylL* (RTY00289) in *S. typhi*, *xylM* (RKP06292), *xylN* (RKP06290), *xylA2* (RKP07393), *xynT* (RKP09414) and *xynB* (RKP04766) in *K. pneumoniae*.

Table 2 (*continued*)

Genome	Operon		Site	Pos.	Score	Dir.	CRP site	Pos.	Score
K. pneumoniae	UG	*xylFGHR*	CAATTACGTTATTTCAC-(4)-CcATTgaGAgAaTcCAC	−308	9.04	⇐	AAAcGTGcgCcAGcTCgCAaaa	−270	4.07
		xylFGHR	acGAAATtTCaTAAcgG-(4)-aTaAgAaccgGTAATTG	−145	7.84	⇒	AAATacGATCgccgTCAtAaTT	−177	3.85
		xylA2/xynTB	GaGcAtTtTCtcAATcG-(4)-GTGAAATAACGTAATTG	−110	9.06	⇒	AttTGcGAgaTgGcTCACATTT	−209	4.75
		xynTB/xylA2	GTaAAAaAACaTAATTa-(4)-aTttAAaAACGTAATaG	−104	9.34	⇒	cttTGcttcgagGATCACAgaa	−126	3.12
		yicJI	aacAAgaAACGTAATct-(4)-acGAgATATCGcAATTc	−113	7.44	⇒	AttccctAcCgcGATCACAgaa	−135	3.05
		xylMN	aacAgAaAACaTAAacc-(4)-GcGAAAaATgacAATaG	−95	7.42	⇒	AtgTGTGAcaactttTCtCAgTT	−117	3.85
Y. pestis	CG	*xylAB/xylFGHR*	aatAAtTcTCtgAATTt-(4)-GTGAAATAACGTAATTG	−198	9.17	⇒	ttATGaGATCTAcAcCACAaTT	−220	4.55
		xylFGHR/xylAB	accAAAaAACaTAATTG-(4)-aTGAAAatctGTAATTG	−128	9.89	⇒	AAAcaTGATCgttATCAtAaaa	−160	4.09
H. influenzae	CG	*HI1112-3/HI1111-09*	GcaAAtaATCaacATaG-(4)-aTtAAATAACaTAATTG	−98	9.20	⇒	AAcTGTGATCcAcgcCACAgTT	−120	4.18
		HI1111-09/HI1112-3	GTGAAAaAACGTAATaa-(4)-GaaAgATtTtaTAATTG	−113	9.29	⇒			
		HI1106					tAtTaTaActTAaATagCAaTT	−42	3.45
A. actinomycetem-comitans	UG	*xylAB/xylFGH*	aacAAtaATCtTAAaTc-(4)-GcGAAATAACGTAATTG	−96	8.96	⇒	tttTGTGAcCcAGtcCACAaTa	−118	4.59
		xylFGH/xylAB	GaGAAAaAACGTAATaG-(4)-GgaAgATcTgaTAATTG	−139	9.09	⇒			
		xylR					tttTtTGAgtgAaATCACAgag	−93	4.12
The XylR half-site consensus			**RtgAAAWAwCrTAATTG**						

porter. A similar situation was experimentally studied in *E. coli* [5].

The regulators and, consequently, the mode of the transcriptional regulation of the arabinose and xylose utilization systems are different in the *Bacillus/Clostridium* group and in gamma *Proteobacteria*. The principal components required to metabolize the respective monosaccharide, as one can expect, are the same, although not necessarily closely homologous. On the other hand, additional members of the regulons were found in many cases.

The AraC regulon is well studied in *E. coli*. Among gamma *Proteobacteria* it is present in enterics only. It is not surprising that the structure of the arabinose regulon in enteric bacteria is very similar to that of *E. coli*, although some gene deletions and operon rearrangements were found. In Gram-positive bacteria, the arabinose regulons are more diverse and include several operons which probably serve to utilize the oligosaccharides containing arabinose.

The xylose regulons in both considered groups of bacteria demonstrate some similar features, despite the difference in the regulator structures. In addition to the preserved core of the regulon for the utilization of xylose itself, a number of operons were found comprising putative transporters and sugar hydrolases. Thus, the experimentally studied *xylPQ* operon of *L. pentosus* [23] seems to represent a quite common phenomenon. The operons having similar structures and presumably functions appear to be under xylose regulation as well, both in Gram-positive and Gram-negative bacteria. It should be mentioned that the genes for the xylose utilization enzymes are found also in the recently sequenced and unfinished genomes of the *Rhizobiaceae* group of alpha *Proteobacteria*. In the *Sinorhizobium meliloti* and *Agrobacterium tumefaciens* genomes, the *xylA* and *xylB* genes form a putative operon with a LacI family regulator, which might function as XylR. The putative xylose regulons of the *Rhizobiaceae* group species will be described elsewhere.

The gene expression of the ribose, xylose and arabinose regulons is modulated by the glucose availability both in the *Bacillus/Clostridium* group and in gamma *Proteobacteria*. However, the mechanisms of such catabolite repression are quite different in the two taxonomic groups. Conserved relative positioning of XylR and CRP in gamma *Proteobacteria* suggests close functional interaction between these regulators.

Thus, the comparative analysis allowed us not only to find new genes likely to be involved in utilization of pentoses, but to consider the evolution of the pentose regulons, to describe common features of these regulons at large evolutionary distances, to determine the binding signals, and even to make predictions about the mechanism of regulation by the responsible transcription factors.

Acknowledgements

This study was partially supported by Grants from the Russian Fund of Basic Research (99-04-48247 and 00-15-99362), the Russian State Scientific Program 'Human Genome', INTAS (99-1476) and HHMI (55000309). We are grateful to D.A. Rodionov for valuable discussions and the CRP recognition profile.

References

[1] Ogden, S., Haggerty, D., Stoner, C.M., Kolodrubetz, D. and Schleif, R. (1980) The *Escherichia coli* L-arabinose operon: binding sites of the regulatory proteins and a mechanism of positive and negative regulation. Proc. Natl. Acad. Sci. USA 77, 3346–3350.

[2] Stoner, C. and Schleif, R. (1983) The *araE* low affinity L-arabinose transport promoter. Cloning, sequence, transcription start site and DNA binding sites of regulatory proteins. J. Mol. Biol. 171, 369–381.

[3] Hendrickson, W., Stoner, C. and Schleif, R. (1990) Characterization of the *Escherichia coli ara*FGH and *ara*J promoters. J. Mol. Biol. 215, 497–510.

[4] Song, S. and Park, C. (1997) Organization and regulation of the D-xylose operons in *Escherichia coli* K-12: XylR acts as a transcriptional activator. J. Bacteriol. 179, 7025–7032.

[5] Song, S. and Park, C. (1998) Utilization of D-ribose through D-xylose transporter. FEMS Microbiol. Lett. 163, 255–261.

[6] Mauzy, C.A. and Hermodson, M.A. (1992) Structural and functional analyses of the repressor, RbsR, of the ribose operon of *Escherichia coli*. Protein Sci. 1, 831–842.

[7] Schleif, R. (2000) Regulation of the L-arabinose operon of *Escherichia coli*. Trends Genet. 16, 559–565.

[8] Lee, N., Francklyn, C. and Hamilton, E.P. (1987) Arabinose-induced binding of AraC protein to *araI2* activates the *araBAD* operon promoter. Proc. Natl. Acad. Sci. USA 84, 8814–8818.

[9] Brunelle, A. and Schleif, R. (1989) Determining residue-base interactions between AraC protein and *ara*I DNA. J. Mol. Biol. 209, 607–622.

[10] Seabold, R.R. and Schleif, R.F. (1998) Apo-AraC actively seeks to loop. J. Mol. Biol. 278, 529–538.

[11] Dunn, T.M., Hahn, S., Ogden, S. and Schleif, R.F. (1984) An operator at −280 base pairs that is required for repression of *araBAD* operon promoter: addition of DNA helical turns between the operator and promoter cyclically hinders repression. Proc. Natl. Acad. Sci. USA 81, 5017–5020.

[12] Martin, K., Huo, L. and Schleif, R.F. (1986) The DNA loop model for *ara* repression: AraC protein occupies the proposed loop sites in vivo and repression-negative mutations lie in these same sites. Proc. Natl. Acad. Sci. USA 83, 3654–3658.

[13] Clarke, P., Lee, J.H., Burke, K. and Wilcox, G. (1992) Mutations in the *araC* gene of *Salmonella typhimurium* LT2 which affect both activator and auto-regulatory functions of the AraC protein. Gene 117, 31–37.

[14] Warren, R.A.J. (1996) Microbial hydrolysis of polysaccharides. Annu. Rev. Microbiol. 50, 183–212.

[15] Rodionov, D.A., Mironov, A.A. and Gelfand, M.S. (2001) Transcriptional regulation of pentose utilization systems in the *Bacillus*/*Clostridium* group of bacteria. In press.

[16] 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.

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

[18] Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M.D., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Hermjakob, H., Hulo, N., Jonassen, I., Kahn, D., Kanapin, A., Karavidopoulou, Y., Lopez, R., Marx, B., Mulder, N.J., Oinn, T.M., Pagni, M. and Servant, F. (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res. 29, 37–40.

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

[20] 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.

[21] 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.

[22] He, B., Choi, K.Y. and Zalkin, H. (1993) Regulation of *Escherichia coli glnB*, *prsA* and *speA* by the purine repressor. J. Bacteriol. 175, 3598–3606.

[23] 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.

[24] Poolman, B., Knol, J., van der Does, C., Henderson, P.J.F., Liang, W.-J., Leblanc, G., Pourcher, T. and Mus-Veteau, I. (1996) Cation and sugar selectivity determinants in a novel family of transport proteins. Mol. Microbiol. 19, 911–922.

[25] Henrissat, B. and Davies, G. (1997) Structural and sequence-based classification of glycoside hydrolases. Curr. Opin. Struct. Biol. 7, 637–644.

[26] Moracci, M., Cobucci Ponzano, B., Trincone, A., Fusco, S., De Rosa, M., van der Oost, J., Sensen, C.W., Charlebois, R.L. and Rossi, M. (2000) Identification and molecular characterization of the first α-xylosidase from an archaeon. J. Biol. Chem. 275, 22082–22089.

[27] Erlandson, K.A., Park, J.-H., El Khal, W., Kao, H.-H., Basaran, P., Brydges, S. and Batt, C.A. (2000) Dissolution of xylose metabolism in *Lactococcus lactis*. Appl. Environ. Microbiol. 66, 3974–3980.