Molecular Microbiology (2000) **38**(4), 673–683

Transcriptional regulation of transport and utilization systems for hexuronides, hexuronates and hexonates in gamma purple bacteria

Dmitry A. Rodionov,* Andrey A. Mironov, Alexandra B. Rakhmaninova and Mikhail S. Gelfand
State Scientific Center GosNIIGenetika, Moscow, 113545, Russia.

Summary

The comparative approach is a powerful tool for the analysis of gene regulation in bacterial genomes. It can be applied to the analysis of regulons that have been studied experimentally as well as that of regulons for which no known regulatory sites are available. It is assumed that the set of co-regulated genes and the regulatory signal itself are conserved in related genomes. Here, we use genomic comparisons to study the regulation of transport and utilization systems for sugar acids in gamma purple bacteria *Escherichia coli*, *Salmonella typhi*, *Klebsiella pneumoniae*, *Yersinia pestis*, *Erwinia chrysanthemi*, *Haemophilus influenzae and Vibrio cholerae*. The variability of the operon structure and the location of the operator sites for the main transcription factors are demonstrated. The common metabolic map is combined with known and predicted regulatory interactions. It includes all known and predicted members of the GntR, UxuR/ExuR, KdgR, UidR and IdnR regulons. Moreover, most members of these regulons seem to be under catabolite repression mediated by CRP. The candidate UxuR/ExuR signal is proposed, the KdgR consensus is extended, and new operators for all transcription factors are identified in all studied genomes. Two new members of the KdgR regulon, a hypothetical ATP-dependent transport system OgtABCD and YjgK protein with unknown function, are detected. The former is likely to be the transport system for the products of pectin degradation, oligogalacturonides.

Introduction

The availability of many complete bacterial genomes allows one to conduct large-scale proteome comparisons with the aim of functional annotation of poorly characterized organisms. Identification of regulatory sites is the next important step in genome annotation. The standard method is to run profile searches using weight matrices trained on samples of experimentally mapped binding sites of transcription factors (Robison *et al*., 1998; Thieffry *et al*., 1998). Indeed, rational choice of positional nucleotide weights leads to reasonable correlation between the computed scores and the binding affinity (Mulligan *et al*., 1984; Berg and von Hippel, 1987; Goodrich *et al*., 1990). However, even for relatively well-studied sites, such as the CRP and SOS boxes of *Escherichia coli*, it is impossible to set thresholds distinguishing reliably between true and non-functional sites. Simultaneous analysis of several related genomes allows one to make reliable predictions even with weak rules. This approach is based on the assumption that regulons (sets of co-regulated genes) are conserved in related genomes. Thus, the true sites occur consistently upstream of orthologous genes, whereas false positives are scattered at random. Of course, this approach can be applied only if orthologous regulators are present in all studied genomes. Other possible complications are drift of the recognition signals and changes in the operon structure. However, this approach does not require that the same recognition rule is applied to all genomes and, thus, changes in the signals can be taken into account. As for the changes in the operon structure, they can be accounted for if the analysis is extended to all genes that can be co-transcribed with the candidate regulon members: it is sufficient to require that the genes are transcribed in the same direction, and that the intergenic spacers are not too long.

Previously, we have applied the comparative approach to the analysis of purine, arginine and aromatic amino acid regulons (Mironov *et al*., 1999), heat shock, SOS and multiple drug resistance regulons of eubacteria, as well as to several archaeal regulons (reviewed by Gelfand, 1999; see also Stojanovic *et al*., 1999; Gelfand *et al*., 2000). Here, we apply the comparative approach to an analysis of the GntR, UxuR/ExuR and KdgR regulons involved in the sugar metabolism in gamma purple bacteria.

Escherichia coli is capable of using hexonates, hexuronates and hexuronides as sources of carbon and energy (Lin, 1996). All these sugar acids are catabolized via enzymes whose expression is regulated by seven

Accepted 20 July, 2000. *For correspondence. E-mail rodionov@genetika.ru; Tel. (+7) 095 315 01 56; Fax (+7) 095 315 05 01.

different transcription factors (for a review, see Peekhaus and Conway, 1998a). The catabolism of gluconate is controlled by GntR. In addition, utilization of idonate, a gluconate predecessor, is subject to positive regulation by IdnR. Transport and catabolism of galacturonate and glucuronate are mediated by genes regulated by UxuR and ExuR. In addition, genes responsible for the utilization of glucuronides are controlled by UidR. KdgR regulates the genes for the catabolism of KDG (2-keto-3-deoxygluconate) and its predecessors. Finally, many operons in these pathways are regulated by CRP.

Thus, the common function of these regulons is to supply phosphogluconate and KDPG for further utilization by the pentose phosphate and Entner–Doudoroff pathways. We consider the complete genomes of *E. coli* and *Haemophilus influenzae and* the unfinished genomes of *Salmonella typhi*, *Yersinia pestis and Vibrio cholerae*. We start with the identification of all relevant orthologous genes in the studied genomes. Then, we use several variants of comparative analysis, dependent on the availability of experimentally determined regulatory sites. We have constructed the KdgR box profile using the training set of *Erwinia chrysanthemi* KdgR binding sites and use it to describe the KdgR regulon in *Y. pestis and Klebsiella pneumoniae* and to predict the transport system for oligogalacturonates in these genomes. We have used the upstream regions of UxuR/ExuR-regulated genes to determine the previously unknown UxuR/ExuR signal and used the constructed profile to find UxuR/ExuR binding sites in other genomes. Finally, we have used the GntR and CRP box profiles derived from experimentally mapped *E. coli* sites to describe the respective regulons in all genomes. The combination of metabolic maps with regulatory networks for three regulons shows the differences in the structure of metabolic pathways in related gamma purple bacteria (Fig. 1).

Results and discussion

GntR regulon

The catabolism of gluconate via the Entner–Doudoroff (ED) pathway in *E. coli* is controlled by the transcription factor GntR (Fig. 1). Gluconate is a true inducer of GntR. The GntR regulon consists of three operons, *gntKU*, *gntT* and *edd-eda* (Fig. 2A), containing *gntT* and *gntU*, which encode high- and low-affinity gluconate transporters, respectively, a thermoresistant gluconokinase *gntK* and two genes from the ED pathway, *edd and eda* (Egan *et al*., 1992; Izu *et al*., 1997; Peekhaus *et al*., 1998b).

The third gluconate transporter gene, *gntP*, is not regulated by GntR (Klemm *et al*., 1996). Enzymes for utilization of the gluconate predecessor, idonate, are encoded by adjacent *idnDOTR* and *gntV* operons that are under positive regulation by IdnR (Bausch *et al*., 1998). The idonate transport protein IdnT, the idonate regulator IdnR and the thermosensitive kinase GntV are homologous to the gluconate transporter GntT, the gluconate regulator GntR and the kinase GntK respectively.

S. typhi has a similar layout of orthologous genes, whereas the operon structure in *Y. pestis and V. cholerae* is different. *H. influenzae* has no genes of the gluconate regulon, nor the regulator itself.

The GntR box profile was constructed by applying SIGNALX (see *Experimental procedures*) to a set of known GntR-regulated genes from *E. coli* and related genomes (Table 1A). The constructed palindromic profile with consensus sequence AAATGTTACCGGTAACATTA assigns a high *Z*-score to the known GntR binding sites upstream of the *gntT* gene in *E. coli.* The positions of the candidate GntR boxes in upstream regions of GntR-regulated genes are shown in Fig. 2A. In many cases, the candidate sites occur in pairs, similar to the known sites of the *gntT* gene in *E. coli*. The distances between the major sites upstream of *gntK* are the same (13–14 bp) in different genomes. It may reflect possible co-operative interactions of GntR dimer pairs.

In *Y. pestis*, the genes of the ED pathway, *edd and eda*, are in different operons that have no candidate GntR boxes. Furthermore, the available part of the *Y. pestis* genome has no orthologue for the second gluconate transporter gene, *gntT*. Thus, the GntR regulon in *Y. pestis* seems to consist only of *gntK* and *gntU*.

In *V. cholerae*, two adjacent operons *gntK–edd* and *gntU–eda* have a common regulatory region and seem to be regulated by GntR via two closely located candidate sites. Moreover, in this genome, the *gntR* gene probably autoregulates itself, whereas in the other genomes, it does not.

There is a candidate GntR box in the common regulatory region of the adjacent *idnDOTR* and *gntV* operons involved in the idonate utilization by *E. coli*. In *Y. pestis*, there is a similar GntR box in the upstream region of the *idnOV* operon. In *S. typhi*, the corresponding operon is interrupted by an integrase gene, and there are no upstream GntR boxes. The regulatory protein IdnR encoded by the *idnDOTR* operon is closely homologous to GntR (53% identity) and, thus, the appearance of the GntR boxes upstream of *idnDOTR* can reflect the similarity of signals recognized by these two transcription factors.

Finally, there is a strong GntR box upstream of a *Y. pestis* sugar dehydrogenase gene denoted *sdxX* in Fig. 2 that has no orthologues in other genomes.

UxuR/ExuR regulon

Transport and catabolism of galacturonate and glucuronate are mediated by genes from the UxuR/ExuR

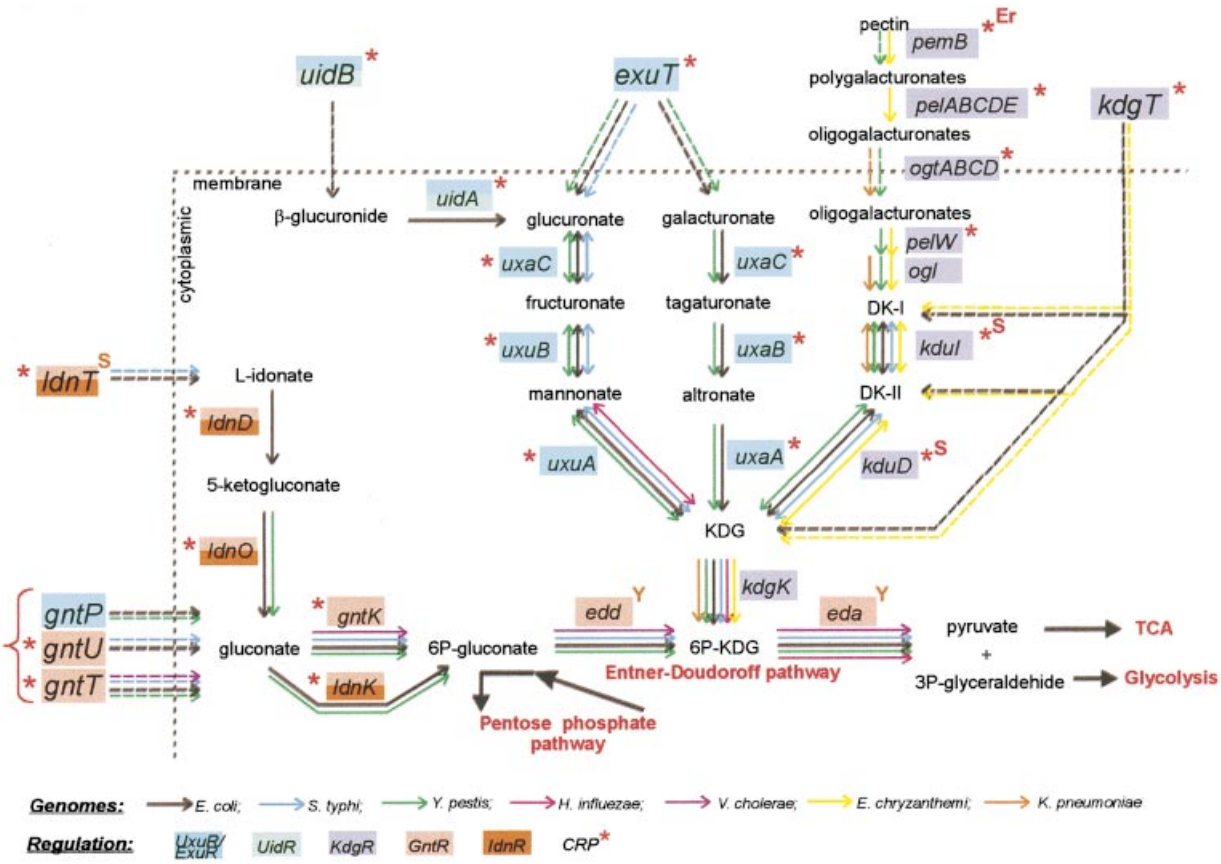


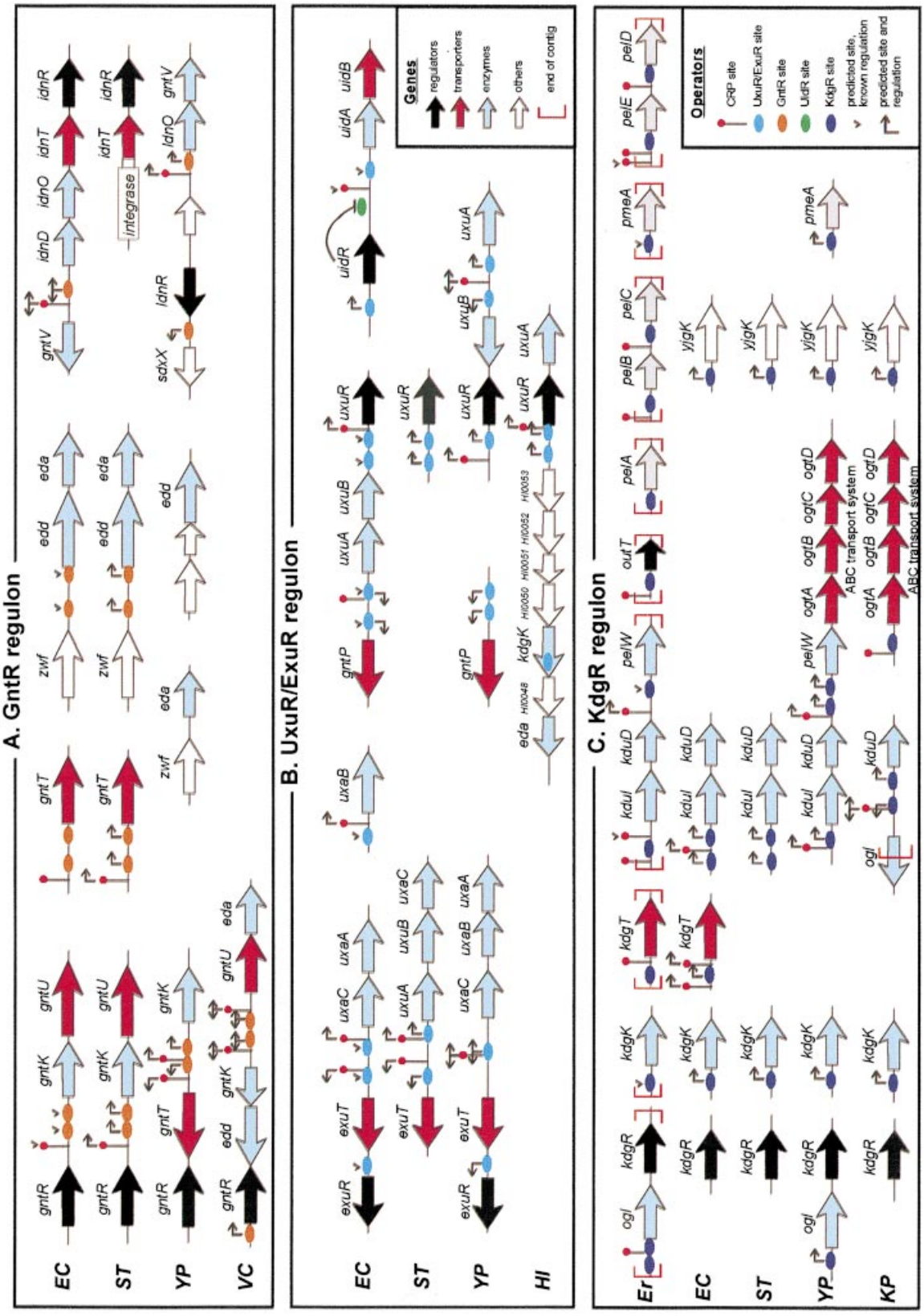
Fig. 1. Transcriptional regulation of the transport and utilization systems for hexuronides, hexuronates and hexonates in gamma purple bacteria. Enzymes and transporters are shown by solid and dashed arrows respectively. Gene regulation is shown by the background colour or asterisk. Superscripts denote loss of regulation in *Salmonella typhi* (S), *Yersinia pestis* (Y), *Erwinia chrysanthemi* (Er). Only the KdgR regulon is shown for the least complete genome of *Klebsiella pneumoniae*.

regulon. Two transcription factors of this regulon, ExuR and UxuR, are highly similar (49% identity) and capable of cross-talk (Ritzenthaler *et al*., 1985). ExuR regulates expression of the *uxaCA*, *uxaB* and *exuT* operons involved in galacturonate catabolism. However, ExuT and UxaC allow for the transport of both galacturonate and glucuronate with consequent isomerization to fructuronate and tagaturonate respectively. Another transcription factor, UxuR, regulates the *uxuAB* operon specifically involved in the catabolism of glucuronate. The repressors ExuR and UxuR apparently act together to mediate regulation of the UxuR regulon. The transport of glucuronides and their subsequent interconversion to the glucuronate in *E. coli* is mediated by the genes of the *uidAB* operon, which is under negative regulation by UidR and UxuR (Blanco, 1987). Finally, the genes *exuR* and *uxuR* are autoregulated by the repressors themselves. All the genes mentioned above are functionally linked to the utilization of various hexuronates, glucuronates or galacturonates (Fig. 1). No binding sites for either UxuR or ExuR have been mapped experimentally.

The operon structures of the UxuR/ExuR regulon genes are shown in Fig. 2B. *S. typhi* and *H. influenzae* have no *uxaA* and *uxaB* genes for the utilization of galacturonates, nor the regulatory gene *exuR*. *S. typhi* has *uxuCBA*, *exuT* and *uxuR* operons, and *H. influenzae* has only the *uxuRA* operon for the utilization of hexuronates. *Y. pestis* has orthologues of all genes of the *E. coli* UxuR/ExuR regulon, but the operon structures are different. Finally, *V. cholerae* has no genes of the UxuR/ExuR regulon.

SIGNALX was applied to the sample of upstream regions of all UxuR/ExuR-regulated genes and their orthologues. This resulted in the identification of palindromic sites with consensus AAATTGGTATACCAATTT and construction of the profile that is likely to represent the UxuR/ExuR signal (Table 1B). In *S. typhi*, *Y. pestis* and *H. influenzae*, all genes orthologous to the genes from the *E. coli* UxuR/ExuR regulon have candidate UxuR/ExuR boxes (Fig. 2B). In addition, we have observed the consistent occurrence of UxuR/ExuR boxes upstream of the *gntP* genes encoding gluconate transporters in *E. coli* and *Y. pestis*.

Another likely member of the UxuR/ExuR regulon is the


A. GntR regulon
B. UxuR/ExuR regulon
C. KdgR regulon
EC
ST
YP
VC
HI
Er
KP
Genes
regulators
transporters
enzymes
others
end of contig
Operators
CRP site
UxuR/ExuR site
GntR site
UidR site
KdgR site
predicted site; known regulation
predicted site and regulation
ABC transport system

uidAB operon responsible for the transport and catabolism of glucuronides in *E. coli*. Indeed, this metabolic pathway merges with the pathway for the utilization of glucuronates (Fig. 1). The *uidAB* operon is controlled by two transcriptional factors, UxuR and UidR, the latter being encoded by the upstream gene of the operon. There are candidate UxuR/ExuR boxes upstream of the *uidA* and *uidR* genes.

There are no obvious differences between the sites upstream of genes regulated by both UxuR and ExuR and those upstream of genes regulated by ExuR only. There is also no specific candidate UxuR signal upstream of the former set of genes. Given the high similarity of the factors, it is likely that they bind the same sites, albeit with different affinity.

KdgR regulon

The two hexuronate pathways lead to the common intermediate 2-keto-3-deoxygluconate (KDG). KDG is phosphorylated by the KDG kinase to form KDPG, the substrate of Eda. The genes for the catabolism of KDG and its predecessors are regulated by transcription factor KdgR.

The KdgR regulon has been studied in detail only in the plant pathogen *E. chrysanthemi* that degrades and uses pectin (Hugouvieux-Cotte-Pattat *et al*., 1996). In *E. chrysanthemi*, KdgR regulates almost all genes involved in pectin degradation (Fig. 1). Extracellular pectate lyases, produced by *E. chrysanthemi* and many other bacteria, including *Yersinia pseudotuberculosis*, convert pectin into oligogalacturonate. Oligogalacturonate can be metabolized into KDG by cytoplasmic enzymes encoded by the *ogl and kduID* operons. Moreover, KDG and its predecessors DKI and DKII can enter the cell via the KdgT transport system. In *E. chrysanthemi*, KdgR also controls the majority of the pectinase genes (*pelA*, *B*, *C*, *D* and *E*) and the *out* genes involved in pectinase secretion (Nasser *et al*., 1994; Rouanet *et al*., 1999).

The related enterobacteria *E. coli*, *S. typhi*, *K. pneumoniae* and *Y. pestis* also have the *kdgR* gene. The operon structures of the orthologues of all *E. chrysanthemi* KdgR-regulated genes are shown in Fig. 2C. The studied genomes do not have genes for the pectinases or the pectinase secretion system. The three genes for the catabolism of KDG and its predecessors, *kdgK*, *kduI* and *kduD*, are present in the genomes of all studied enterobacteria; the absence of *kduI* in the *K. pneumoniae* genome is probably caused by the unfinished state of this genome. The *kdgT* gene, encoding the transporter of KDG, is present only in *E. coli*. The gene organization of *Y. pestis* is the closest to that of *E. chrysanthemi*, as there are two genes for the utilization of oligogalacturonides, *ogl* and *pelW*, *and* the pectinesterase gene *pmeA*.

SIGNALX was applied to the set of upstream regions of all known KdgR-regulated genes from *E. chrysanthemi* (Table 1C). The derived consensus sequence of the palindromic KdgR box is AAATGAAACAnTGTTTCATTT. It is much wider than described previously (Nasser *et al*., 1994). Using the constructed KdgR profile, we reconstructed the KdgR regulon in other bacteria. In *E. coli*, it consists of the *kdgK*, *kdgT* and *kduID* operons. The KdgR regulon of *S. typhi* is similar to that of *E. coli*, but there is no *kdgT* gene. All *Y. pestis* genes orthologous to the KdgR-regulated genes from *E. chrysanthemi* have strong KdgR boxes. Immediately downstream of the *pelW* gene, there are four genes encoding a hypothetical ATP-binding cassette transport system (ABC system) (Fig. 2C). Short intergenic distances make it highly likely that they form a single operon with *pelW*. The first two genes encode the transmembrane components of the ABC transport system, whereas the third and the fourth genes encode the ATP- and substrate-binding components respectively. These proteins are homologous to various disaccharide ABC transporters from the MAlF-G-E-K subfamily (Fig. 3). Oligogalacturonates are the substrate for two cytoplasmic oligogalacturonate lyases, PelW and Ogl (Shevchik *et al*., 1999). It is well known that functionally coupled genes are often located in close proximity to each other or organized in operons (Bork and Koonin, 1998; Overbeek *et al*., 1999), so the function of this transport cassette is likely be linked with *pelW* (Fig. 1). We propose that the function of these proteins is the active transport of oligogalacturonates in the cytoplasm and name this system OgtA-B-C-D (oligogalacturonate transport). In *Y. pestis*, the single *pelW–ogtABCD* operon has a strong KdgR box. In *E. chrysanthemi*, the sequence downstream of *pelW* is unavailable. Orthologues of *ogtABCD* were found in the *K. pneumoniae* genome (Fig. 2C) and provide additional support for the above prediction. Indeed, despite a different gene arrangement (no *pelW* upstream of *ogtABCD*), there is a strong KdgR box upstream of *ogtABCD* in *K. pneumoniae.*

Finally, the upstream region of *yjgK* contains strong KdgR boxes in the *E. coli*, *S. typhi*, *K. pneumoniae* and *Y. pestis* genomes. Multiple alignment of *yjgK* upstream ‘regions does not contain any conserved regions and, thus, these boxes are likely to be functional. The functional relevance of *yjgK* for the KdgR regulon is unclear, because this gene has no significant homologues in the other complete genomes, nor in the sequence databases.

Fig. 2. Operon structures and regulatory sites of the GntR, UxuR/ExuR and KdgR regulons in *Escherichia coli* (EC), *Salmonella typhi* (ST), *Klebsiella pneumoniae* (KP), *Yersinia pestis* (YP), *Erwinia chrysanthemi* (Er), *Haemophilus influenzae* (HI) and *Vibrio cholerae* (VC).

Table 1. Site scores and positions relative to the operon start of (A) GntR, (B) UxuR/ExuR, (C) KdgR and (D) CRP.

Operons	Position	Score	Binding site(s)	*
A. GntR regulon				
E. coli				
gntKU	−99	5.84	**cAATGTTACCGaTAACAgTT**	R[a]
	−86	5.3	AAcaGTTACCcGTAACATTT	
gntT	−179	5.69	**AgATGTTACCcGTAtCATTc**	S[a]
	−38	4.42	TgAcGTTACCcaTAACAaaT	S[a]
edd-eda	−211	6	**TtATtTTACCGGTAACATgA**	R[a]
	11	4.65	AAtTGTTACgcGTAACAaaT	
idnDOTR/gntV	−116	4.82	TcAcGTTAtgcGTAACATag	R[b]
S. typhi				
gntKU	−101	5.84	**cAATGTTACCGaTAACAgTT**	
	−88	4.62	AAcaGTTACCcGTAACAaaT	
gntT	−176	5.78	**AtATGTTACCcGTAtCATTc**	
	−35	4.36	TgAcGTTACCcaTAACAaag	
edd-eda	−213	5.94	**gtATtTTACCGGTAACATgA**	
	9	4.48	AttTGTTACgcGTAACAcag	
Y. pestis				
gntT/gntK	−106	6.05	**ccATGTTACCGGTAtCATgA**	
	−92	5.48	TcATGaTACCGGTAACAaaT	
idnO-gntV	−23	4.39	AtATaTTAtCGcTAtCAggA	
sdsX	−220	5.72	TcAcGTTACCGGTAACATgT	
V. cholerae				
gntU-eda/ *gntK-edd*	–81	5.8	TcATGTTAtgGGTAACATgT	
	−67	5.57	AcATGTTACCGGTAACtTcA	
	Consensus		**saATGTTACCGGTAACATtw**	
B. UxuR/ExuR regulon				
E. coli				
exuR	−33	5.26	**AAAgTGGTATAaCAAaTa**	R[c]
uxaB	−194	5.24	**AAtgTGtTcTACCAcTTT**	R[c]
uxuAB/gntP	−181	5.24	**gAtgTGGTtaACCAATTT**	R[d]
	−287	4.5	AAATTGGTcaACCAtTgT	
uxuR	−174	5.06	**tAtTTGGTtgACCAgTTT**	R[d]
exuT/uxaCA	−130	4.28	**AAAgTtGTATgaCAAgTT**	R[c]
	−255	3.82	AatTTttTtaACtAcgTT	
uidR	−128	5.19	ctATTGGTtaACCAATTT	
uidAB	−61	4.14	gAATTGGTtaACtAATca	R[e], R[c]
S. typhi				
exuT/uxuABC	−274	4.42	AAATTGGctagCCAAaTc	
	−87	5.07	**tgATTGGTcaACCAATTT**	
uxuR	−105	4.57	ttgTTGGTtgACCAgTTT	
	−230	5.25	**AtATTGGTAaACCAATaT**	
Y. pestis				
uxuB	−92	4.99	**AAAcTGGTcTAaCAAcTT**	
uxuA	−241	4.44	AAATcGGTATAaCAAcaT	
uxuR	−80	4.66	**AAgTTGtTtgACCAcaTg**	
exuR	−13	4.93	**tAAgTGGTATAaCAAaTc**	
gntP	−229	4.78	AAcTTGGTtaACCAATac	
	−209	4.54	cAtTTGGTtatCCAcTTa	
uxaCBA/exuT	−225	3.72	AAcTTGtcATACagATTT	
H. influenzae				
uxuRA	−23	5.47	**AAATTGGaATACCAATTT**	
	−73	4.39	tttTTGGaATACCAAaTa	
Consensus			**AAATTGGTATACCAATTT**	
C. KdgR regulon				
E. chrysanthemi				
ogl	−150	5.15	**AAATGAAAgAATGTTTtATaa**	S[f]
	−108	5.63	**AAATGAAACgTTGTTTCtaca**	S[f]
kduID	−182	6.05	**AAATaAAACATTaTTTCATTT**	S[f]
pelW	−57	5.42	**AAtcaAAACAATGTTTCtaTT**	R[f]
kdgK	−98	5.64	**AAATaAAACATcGTTTCATcg**	S[f]
kdgT	−56	6.07	**AAAaGAAACATTGTTTCATTT**	S[f]
outT	−138	5.83	**tAATGAAACggTGTTTtATTa**	S[f]
pelA	−215	5.52	**AttTaAAACATcGTTTCATTa**	S[f]
pelB	−171	5.72	**tAATGAAAtggcaTTTCAaTT**	S[f]
pelC	−232	5.44	**tAATGAAAttAcGTTTCAacT**	S[f]
pelE	−81	5.59	**AgATGAAAtggTaTTTCgTTT**	S[f]

Table 1. *continued*

Operons	Position	Score	Binding site(s)	*
pelD	−79	5.21	**ggAcaAAAtggcGTTTCATTT**	S[f]
pemB	−126	5.17	**AAATGAAACgcaGgTTtATTT**	R[f]
E. coli				
kduID	−129	6.16	AAATGAAACATTGTTTtATTT	
	−62	5.34	AAAcGAAACAgTGTTTCAcTa	
yjgK	−27	6.01	AAATGAAACgTTGTTTtAaTT	
kdgT	−132	5.93	AAATaAAACAgcGTTTCAaTT	
kdgK	132	5.11	tAATGgAACAcTGTTTtAaTa	
S. typhi				
kduID	−85	6.15	AAATGAAACgTTGTTTtATTT	
	−20	5.2	AAtcaAAACAgTGTTTtgaTT	
yjgK	−33	5.62	AAATaAAACgcTGTTTtAacT	
kdgK	−62	4.59	tAATGgAcCgATGTTTtAaTa	
K. pneumoniae				
kduD/ogl	−361	5.64	AAATGAAACATcGTTTtAaaT	
	−293	4.95	AAAcGgAACTcTGTTTtATTT	
ogt*A*BCD	−216	5.3	AAtTaAAACggTGTTTtATaa	
kdgK	222	5.18	tAATGgAACgATGTTTtAaTa	
yjgK	−30	5.59	AAATGAAAtgcTGTTTtATaT	
Y. pestis				
kduID	−202	5.92	tAATaAAACATcaTTTCATTT	
pelW-ogtABCD	−153	4.65	AAtcaAAACAAcGTTcCgacT	
	−77	5	gAtTGAnACgATGTTTCtaTT	
ogl	−28	5.81	AAATGAAACATTGTTTCtaTa	
yjgK	−42	5.84	tAATaAAACAgcaTTTCATTT	
kdgK	−196	4.94	AttTaAAACAccGTTTtAaTc	
pmeA	−150	5.5	AAATGgAAtggcGTTTCATTT	
Consensus			**AAATGAAACAnTGTTTCATTT**	
D. CRP regulation				
E. coli				
uxaB	−144	3.61	AAccaTGATCcgcgcCACAcTT	
uidAB	−180	4.21	tAATGcGATCTAtATCACgcTg	
kduID	−170	3.4	AttcGTGATCgAcActgCAcTT	
uxaCA/exuT	−231	4.15	AAAgGTGAgagccATCACAaaT	
	−109	3.9	tttatTGATCTAacTCACgaaa	
gntT	−237	3.85	tAATaTGAcCaAccTCtCATaa	S[g]
gntKU	−170	4.73	AAATtTGAagTAGcTCACAcTT	R[g]
kdgT	−172	4.62	tttTGTGATCaAttTCAaAaTa	
	−112	3.95	tgATGTGgTtTtGATCACtTTT	
gntV/idnDOTR	−158	4.44	AttTGTGATgaAGATCACgTca	
	−108	3.55	tAAcGTGATgTgccTtgtAaTT	
gntP/uxuAB	−174	3.58	AttgGTtAaCcAcATCACAaga	S[g]
	−89	3.56	ggATGTGAcaTtcATCgCAaca	
UxuR	−36	3.78	AAATtTGATtaAccgCACcTaa	
S. typhi				
ExuT/uxuABC	3.95		AttgGctAgCcAaATCACAaaT	
	−222	4.05	ttATtTGATCTgcgTCAatTTT	
gntKU	−172	4.48	AAcTtTGAagTAGcTCACAcTT	
gntT	−176	3.75	AtATGTtAcCcgtATCAttcTT	
Y. pestis				
uxuB/uxuA	−280	3.8	ttggtTGATCcActTCACgaaT	
uxuR	−101	3.56	AAcTGTtAcCTAccTCtaATaT	
uxaCBA	−206	4.04	AAAgtTGAgtgAcATCACAaaa	
gntT/gntK	−70	4.18	AgATGTGActTttATCACAaca	
	−94	3.71	tAtcaTGATaccGgTaACAaaT	
idnO-gntV	−313	4.59	tAtaGTGATCTgcATCACAgTT	
kduID	−224	3.71	AAtcaTGAcgTgccTCAaAaaa	
pelW-ogtABCD	−279	3.66	tAtTGTGAcaTtcAaCtCAaaa	
∼*pmeA*	−193	3.77	AAATtTGATtTttATCAtAgag	
E. chrysanthemi				
kduID	−284	4.23	ttgTGTGAaCaAGgTaACAcaa	S[g]
	−166	3.49	cAtTtTtATtgAatTCACATcT	
pelW	−307	3.81	ttATGTGAaCggttaCACcaTT	
	−240	4.43	tttTGTGATCgAaggCACAaTa	
kdgT	−37	3.45	ttgTtTGcaagcGATCACtTTT	S[g]
ogl	−130	2.95	AAATaaaAcCacGATCACggaa	S[g]
outT	−179	4.34	AAtTtTGAgCctGgTCgCAaaa	

Table 1. *continued*

Operons	Position	Score	Binding site(s)	*
pelB	−217	3.81	tAccGTGAgCctGcTCAaAacT	S[g]
pelC	−277	3.91	AAAaGTGAcgcctgTCAaAaTT	S[g]
pelE	−236	3.96	AtAatTGATtTAaATCAtAaaa	
	−140	4.05	cAtcGTGAcaaAGtTCACAaaa	S[g]
pelD	−139	4.31	AAAcGaGATtTtGATCACAaaa	S[g]
H. influenzae				
uxuRA	−5	4.94	tttTGTGAgCcAtATCACAaaa	
V. cholerae				
gntU-eda/ gntK-edd	−137	3.83	AAActTGcgCgtGATCgCATTT	
	−45	4.41	AAATGTGAgCTAtAaCACAaac	
K. pneumoniae				
ogtABCD	3.75		cgtTtcGActTtGATCACATTT	
Consensus			**wwwTGTGAtyyrgwTCACTtWt**	

The divergently arranged operons are separated by a slash. Sites marked in bold have been included in the learning set. The last column (*) represents experimental data about regulation: S, known binding sites; R, known regulation, binding site not known.Regulons: **a.** GntR; **b.** IdnR; **c.** UxuR/ExuR; **d.** ExuR; **e.** UidR; **f.** KdgR; **g.** CRP.

CRP regulation

There are experimental data showing that the expression of *gntKU* and *gntT* from the GntR regulon and *uxuAB* from the UxuR/ExuR regulon is under catabolite repression by CRP (Izu *et al*., 1997; Peekhaus and Conway, 1998b). Moreover, almost all genes from the KdgR regulon are regulated by CRP in *E. chrysanthemi* (Nasser *et al*., 1997).

Scanning of all studied genomes with the CRP profile detected new CRP boxes upstream of some of the genes above described (Table 1D). In order to decrease the prediction noise, we assumed that a candidate box was significant if it was conserved in all genomes containing the corresponding gene.

This analysis demonstrates that the majority of genes for the transport and utilization of sugar acids are under the regulation of CRP (Fig. 1). This provides for consistent CRP regulation of all feeder pathways leading to the ED pathway.

Conclusions

The closely related enterobacteria *E. coli*, *S. typhi*, *Y. pestis* and *E. chrysanthemi* seem to have the most diverse sugar acid catabolic pathways and the most complicated regulatory interactions. In *E. coli*, the genes related to this part of the sugar metabolism are regulated by at least seven different transcription factors. The catabolite repressor CRP is predicted to regulate all target genes, the only exceptions being the gluconate transporter *gntP* and the main genes of the ED pathway, *edd* and *eda*. The three main regulons, GntR, UxuR/ExuR and KdgR, are required for the utilization of gluconates, hexuronates and predecessors of KDG respectively. In addition, *E. coli* has two local regulators, UidR and IdnR, which regulate single operons for the utilization of glucuronides and idonates respectively. The latter systems are unique to *E. coli*. The complicated case of cross-talking UxuR and ExuR factors is preserved only in *Y. pestis*, but the latter has lost the GntR regulation of the ED pathway. *S. typhi* apparently has no UidR, IdnR and ExuR regulons. The KdgR regulon of the plant pathogen *E. chrysanthemi* is the largest and contains an array of genes for the degradation and subsequent utilization of plant pectin. The KdgR regulon of *Y. pestis* is the closest to that of *E. chrysanthemi.* The predicted new member of the KdgR regulon, an ABC transport system *ogtABCD* in *Y. pestis and K. pneumoniae*, is a missing link for the utilization of oligogalacturonates in the metabolic map. We believe that the KdgR regulon of the related *E. chrysanthemi* also contains an as yet unsequenced transport system orthologous to *ogtABCD* that is required for the uptake of the pectin degradation products.

The two genomes of non-enteric gamma purple bacteria, *H. influenzae and V. cholerae*, have fewer genes of the sugar acid regulons. *V. cholerae* has only the GntR regulon, which consists of the transporter, the gluconate kinase and the two genes of the ED pathway, *edd and eda*. The small genome of the obligate pathogen *H. influenzae* has lost almost all genes, retaining only *uxuA*, *kdgK and eda* regulated by the UxuR repressor.

It is noteworthy that regulation is retained in many cases despite the changes in operon structure (Fig. 2). This provides additional validation for the predictions, e.g. in the case of the transporter cassette *ogtABCD*. As the transcription start points are unknown in most cases, nor can they be predicted with any degree of reliability, we cannot analyse in detail the evolution of the operator–promoter (that is, transcription factor–RNA polymerase) interactions.

Fig. 3. Multiple alignments of the predicted oligogalacturonate ABC transporters OgtA-B-C-D from *Y. pestis* and *K. pneumoniae* with homologous proteins: lactose ABC transporter LacF-G-K (Williams *et al.*, 1992) from *Agrobacterium radiobacter* (respective accession numbers in SWISSPROT: P29823, P29824and Q01937); hypothetical ABC transporter YesO-P-Q from *Bacillus subtilis* (respective accession numbers in SWISSPROT/GenBank: CAB12516, O31519and O31520) and probable multiple sugar-binding ABC transporter MsmX from *B. subtilis* (accession number P94360in SWISSPROT).

Enterobacteria occupy variable ecological niches and possess wide capabilities for using various sugar acids via the ED pathway and several feeder pathways. Other gamma purple bacteria have fewer genes for sugar acid catabolism. Using comparative techniques, we have been able to describe the sugar acid regulons in several gamma purple bacteria, to propose new regulatory signals and to predict additional members of these regulons. We

could not find any consistent differences between the UxuR and ExuR binding sites, nor between the sites of orthologous regulators from different species. The resolution of profile methods is insufficient for reliable prediction of the relative binding affinities of regulators to different sites, especially if the training sets are small, as is the case for all regulons considered in the present study. Thus, the final characterization of the considered regulons should be done experimentally. However, we believe that the results presented here not only facilitate these experiments, allowing for direct verification of the specific predictions, but are sufficient to draw general conclusions about the evolution of the metabolic pathways and their transcriptional regulation. Further research will be directed towards complete characterization of regulation of the sugar metabolism in gamma purple bacteria and extension of this analysis to other phylogenetic groups of bacterial genomes.

Experimental procedures

The complete genome sequences of *E. coli* (Blattner *et al*., 1997) and *H. influenzae* (Fleischmann *et al*., 1995), as well as the partial sequences of *E. chrysanthemi*, were downloaded from GenBank (Benson *et al*., 2000). Preliminary sequence data for the *S. typhi*, *K. pneumoniae*, *Y. pestis and V. cholerae* genomes were obtained from The Institute for Genomic Research website (http://www.tigr.org).

Two approaches to the construction of recognition rules for regulatory signals were used. The first requires a set of upstream regions of orthologous genes that are known to be co-regulated in at least one genome. It is based on the assumption that the sets of co-regulated genes in related genomes and the corresponding regulatory signals are conserved. The second approach uses a known set of co-regulated genes from one genome to construct a profile that is subsequently used to analyse other genomes. The existence of the regulatory gene encoding the corresponding transcription factor is a prerequisite for both variants of analysis. We have used the first approach for analysis of the GntR and UxuR/ExuR regulons and the second one for the KdgR regulon.

A simple iterative procedure implemented in program SIGNALX is performed in order to construct a profile from a set of upstream gene fragments (Gelfand *et al*., 2000). Weak palindromes are selected in each region. Each palindrome is compared with all other palindromes, and the palindromes most similar to the initial one, at most one from each region, are used to make a profile. The positional nucleotide weights in this profile are defined as follows (Mironov *et al*., 1999)

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

where $N(b,k)$ is the count of nucleotide b at position k. The score of a candidate site is calculated as the sum of respective positional nucleotide weights:

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

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 (Schneider *et al*., 1986):

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

where $f(i,k)$ is the frequency of nucleotide i at position k of palindromes generating the profile. The best profile is used as the recognition rule. We have repeated the same procedure without requiring the signal to be palindromic and obtained the same results in all cases.

The profile is used to scan all studied genomes. This results in the identification of new candidate boxes. If all orthologues in the analysed genomes have a candidate box with a significant Z-score, then we consider this box to be a putative regulatory site and include the genes of the corresponding operons in the regulon.

The search profile for CRP boxes was kindly provided by Novichkova (2000). Protein alignment was performed using the Smith–Waterman algorithm implemented in the GENOMEEXPLORER program (Mironov *et al*., 2000). Orthologous proteins were defined as bidirectional best hits (Tatusov *et al*., 2000). Distant homologues were identified using PSI-BLAST (Altschul *et al*., 1997). Multiple sequence alignments were constructed using CLUSTALX (Thompson *et al*., 1997). Site recognition was done using GENOMEEXPLORER (Mironov *et al*. 2000).

Acknowledgements

This study was supported by grants from the Merck Genome Research Institute (244), the Russian Fund for Basic Research (99-04-48247 and 00-15-99362), the Russian State Scientific Program 'Human Genome' and INTAS (99-1476). We are grateful to E. Koonin and M. Galperin for useful discussions, and to E. Novichkova for the CRP recognition profile. This study has been done in part during a visit by D.R, A.M. and M.G. to the National Center for Biotechnology Information, National Institutes of Health, USA.

References

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

Blattner, F.R., Plunkett, G., Bloch, C.A., *et al*. (1997) The complete gene sequence of *Escherichia coli* K-12. *Science* **277**: 1453–1474.

Bausch, C., Peekhaus, N., Utz, C., Blais, T., Murray, E., Lowary, T., *et al*. (1998) Sequence analysis of the GntII (subsidiary) system for gluconate metabolism reveals a novel pathway for L-idonic acid catabolism in *Escherichia coli*. *J Bacteriol* **180**: 3704–3710.

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.

Berg, O.G., and von Hippel, P.H. (1987) Selection of DNA binding sites by regulatory proteins. Statistical–mechanical theory and application to operators and promoters. *J Mol Biol* **193**: 723–750.

Blanco, C. (1987) Transcriptional and translational signals of the *uidA* gene in *Escherichia coli* K12. *Mol Gen Genet* **208**: 490–498.

Bork, P., and Koonin, E.V. (1998) Predicting functions from protein sequences – where are the bottlenecks? *Nature Genet* **18**: 313–318.

Egan, S.E., Fliege, R., Tong, S., Shibata, A., Wolf, R.E., and Conway, T. (1992) Molecular characterization of the Entner–Doudoroff pathway in *Escherichia coli*: sequence analysis and localization of promoters for the *edd–eda* operon. *J Bacteriol* **174**: 4638–4646.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., *et al*. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**: 496–512.

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

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.

Goodrich, J.A., Schwartz, M.L., and McClyre, W.R. (1990) Searching for and predicting the activity of sites for DNA binding proteins: compilation and analysis of the binding sites for *Escherichia coli* integration host factor (IHF). *Nucleic Acids Res* **18**: 4993–5000.

Hugouvieux-Cotte-Pattat, N., Condemine, G., Nasser, W., and Reverchon, S. (1996) Regulation of pectinolysis in *Erwinia chrysanthemi*. *Annu Rev Microbiol* **50**: 213–257.

Izu, H., Adachi, O., and Yamada, M. (1997) Gene organization and transcriptional regulation of the *gntRKU* operon involved in gluconate uptake and catabolism of *Escherichia coli*. *J Mol Biol* **267**: 778–793.

Klemm, P., Tong, S., Nielsen, H., and Conway, T. (1996) The *gntP* gene of *Escherichia coli* involved in gluconate uptake. *J Bacteriol* **178**: 61–67.

Lin, E.C.C. (1996) Dissimilatory pathways for sugars, polyols, and carboxylates. In *Escherichia coli and Salmonella: Cellular and Molecular Biology*, 2nd edn, **Vol. 1**. Neidhardt, F.C., Curtiss, R., III, Ingraham, J.L., Lin, E.C.C., Low, K.B., Magasanik, B., et al. (eds). Washington, DC: American Society for Microbiology Press, pp. 307–342.

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.

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

Mulligan, M.E., Hawley, D.K., Entriken, R., and McClure, W.R. (1984) *Escherichia coli* promoter sequences predict *in vitro* RNA polymerase selectivity. *Nucleic Acids Res* **12**: 789–800.

Nasser, W., Reverchon, S., Condemine, G., and Robert-Baudouy, J. (1994) Specific interactions of *Erwinia chrysanthemi* KdgR repressor with different operators of genes involved in pectinolysis. *J Mol Biol* **18**: 427–440.

Nasser, W., Robert-Baudouy, J., and Reverchon, S. (1997) Antagonistic effect of CRP and KdgR in the transcription control of the *Erwinia chrysanthemi* pectinolysis genes. *Mol Microbiol* **26**: 1071–1082.

Novichkova, E.S. (2000) Comparative Approach to Analysis of Regulation in Complete Genomes: Catabolite Repression in Gamma-proteobacteriae. MSc Thesis, Moscow State Engineering Physics Institute.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D., and Maltsev, N. (1999) The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci USA* **96**: 2896–2901.

Peekhaus, N., and Conway, T. (1998a) What's for dinner? Entner–Doudoroff metabolism in *Escherichia coli*. *J Bacteriol* **180**: 3495–3502.

Peekhaus, N., and Conway, T. (1998b) Positive and negative transcriptional regulation of the *Escherichia coli* gluconate regulon gene *gntT* by GntR and the cyclic AMP (cAMP)–cAMP receptor protein complex. *J Bacteriol* **180**: 1777–1785.

Ritzenthaler, P., Blanco, C., and Mata-Gilsinger, M. (1985) Genetic analysis of *uxuR* and *exuR* genes: evidence for ExuR and UxuR monomer repressors interactions. *Mol Gen Genet* **199**: 507–511.

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.

Rouanet, C., Nomura, K., Tsuyumu, S., and Nasser, W. (1999) Regulation of *pelD* and *pelE*, encoding major alkaline pectate lyases in *Erwinia chrysanthemi*: involvement of the main transcriptional factors. *J Bacteriol* **181**: 5948–5957.

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.

Shevchik, V.E., Condemine, G., Robert-Baudouy, J., and Hugouvieux-Cotte-Pattat, N. (1999) The exopolygalacturonate lyase PelW and the oligogalacturonate lyase Ogl, two cytoplasmic enzymes of pectin catabolism in *Erwinia chrysanthemi* 3937. *J Bacteriol* **181**: 3912–3919.

Stojanovic, N., Florea, L., Riemer, C., Gumucio, D., Slightom, J., Goodman, M., *et al*. (1999) Comparison of five methods for finding conserved sequences in multiple alignments of gene regulatory regions. *Nucleic Acids Res* **27**: 3899–3910.

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.

Thieffry, D., Salgado, H., Huerta, A.M., and Collado-Vides, J. (1998) Prediction of transcriptional regulatory sites in the complete genome sequence of *Escherichia coli* K-12. *Bioinformatics* **14**: 391–400.

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.

Williams, S.G., Greenwood, J.A., and Jones, C.W. (1992) Molecular analysis of the *lac* operon encoding the binding-protein-dependent lactose transport system and beta-galactosidase in *Agrobacterium radiobacter*. *Mol Microbiol* **6**: 1755–1768.