Microbiology (1994), 140, 2349–2354 Printed in Great Britain

Evolutionary relationships between sugar kinases and transcriptional repressors in bacteria

Friedrich Titgemeyer, Jonathan Reizer, Aiala Reizer and Milton H. Saier, Jr

Author for correspondence: Milton H. Saier, Jr. Tel: +1 619 534 4084. Fax: +1 619 534 7108. e-mail: msaier@ucsd.edu

Department of Biology, University of California at San Diego, La Jolla, CA 92093-0116, USA

We have characterized a new family of proteins (the ROK family) which includes six transcriptional repressors for sugar catabolic operons, three sugar kinases, and three unidentified open reading frames. Analyses of the aligned sequences and phylogenetic tree construction allow predictions regarding the functional nature of conserved domains and residues within these proteins as well as the pathway of evolutionary divergence that gave rise to the family.

Keywords: transcription, repressors, sugar kinases, phylogenetic family, evolution

INTRODUCTION

In earlier studies, we reported that a large family of repressor proteins including LacI, GalR and FruR is homologous to a class of bacterial periplasmic sugar-binding receptors (Vartak *et al.*, 1991; Tam & Saier, 1993a). These proteins are homologous only in their C-terminal, sugar-binding regions, and the receptor proteins lack the N-terminal DNA-binding domains that contain identifiable helix-turn-helix motifs (Vartak *et al.*, 1991; Weickert & Adhya, 1992). The periplasmic receptors, on the other hand, possess N-terminal, hydrophobic targeting sequences which are lacking in the cytoplasmic repressors. Thus, it appears that gene splicing and fusion events, resulting in shuffling of protein domains, gave rise to this family of homologous proteins that include cytoplasmic repressors and extracellular receptors.

A second family of evolutionarily related proteins includes periplasmic receptors for aliphatic amino acids and a single sequenced bacterial transcriptional regulatory protein (Tam & Saier, 1993a). The latter protein is the cytoplasmic AmiC repressor of the aliphatic amidase structural gene in *Pseudomonas aeruginosa* (Wilson & Drew, 1991). As for the large family of sugar-binding proteins noted above, the homologous domains of the aliphatic-amino-acid-binding proteins and of AmiC are the ligand-binding domains. AmiC presumably binds aliphatic amines as inducers, and these compounds are structural analogues of the aliphatic amino acids.

Gram-positive bacteria including *Bacillus subtilis*, *Staphylococcus xylosus* and *Lactobacillus pentosus* can utilize D-xylose via an inducible process which feeds directly into the pentose phosphate pathway (Kreuzer *et al.*, 1989; Lokman *et al.*, 1991; Sizemore *et al.*, 1991). The inducibility of the enzymes involved is controlled by repressor proteins, designated XylR, that have recently been shown to be homologous to a glucokinase from *Streptomyces coelicolor* (Angell *et al.*, 1992). The report of Angell *et al.* (1992) led us to consider the possibility that a third family of repressor proteins had co-evolved by common descent with non-DNA-binding proteins, in this case with a family of sugar kinases, via a pathway involving domain shuffling. We have analysed this family of proteins, identified all current protein members and constructed a phylogenetic tree. The family contains three major groups of proteins. The first group includes DNA-binding proteins that also bind sugars (xylose or *N*-acetylglucosamine). The second group includes sugar kinases specific for fructose and glucose, and the third group consists of three functionally uncharacterized ORFs. All of these proteins are of bacterial origin. Multiple alignments revealing the regions of conservation and strongly conserved residues are presented. These analyses represent the first step in the structural and functional characterization of this novel family of proteins.

RESULTS

Table 1 presents a list of the proteins which proved to be homologous to XylR of *B. subtilis*. We designate this family the ROK (repressor, ORF, kinase) family. It includes seven sequenced xylose repressors (XylR; Kreuzer *et al.*, 1989; Sizemore *et al.*, 1991; Lokman *et al.*,

Table 1. Proteins included in the ROK family

Abbreviation	Biological source	Gene	No. of residues	GenBank accession no.	References
XylR(BsuI)*	*Bacillus subtilis*	*xylR*	384	M27248	Kreuzer *et al.* (1989)
XylR(BsuII)	*Bacillus subtilis*	*xylR*	384	A00033	S. Hastrup, unpublished
XylR(Tba)	Thermophilic bacterium	*xylR*	399	L18965	P. P. Dwivedi and others, unpublished
NagC(Eco)	*Escherichia coli*	*nagC*	406	M19284	Plumbridge (1989)
XylR(Sxy)	*Staphylococcus xylosus*	*xylR*	383	X57599	Sizemore *et al.* (1991)
XylR(Lpe)	*Lactobacillus pentosus*	*xylR*	388	M57384	Lokman *et al.* (1991)
GlcK(Sco)	*Streptomyces coelicolor*	*glcK*	317	S26208	Angell *et al.* (1992)
FruK(Zmo)	*Zymomonas mobilis*	*frk*	301	M97296	Zembrzuski *et al.* (1992)
ScrK(Smu)	*Streptococcus mutans*	*scrK*	293	D13175	Sato *et al.* (1993)
Orf309(Eco)	*Escherichia coli*		309	Ecouw89_131†	Blattner *et al.* (1993)
Orf260(Eco)	*Escherichia coli*		260 (fragment)	M64780	Reeder & Schleif (1991)
Orf182(Cpe)	*Clostridium perfringens*		182 (fragment)	M81878	B. Carnard and others, unpublished

* Two additional xylose repressors, from *Bacillus megaterium* and *Bacillus licheniformis* (Rygus *et al.*, 1991; Scheler *et al.*, 1991) have been sequenced, but are not in the database. They exhibit 70% and 49% identity, respectively, to XylR(BsuI).

† Genpept accession number.

Table 2. Percentage identity values and comparison scores for members of the ROK family

Values in parentheses, below the designation of the protein, refer to the number of residues in the intact protein. Values presented in the table that are not in square brackets or parentheses represent percentage identity for segments having the number of compared residues indicated in parentheses. The FASTA program, using the dipeptide mode (ktup = 2) (Pearson & Lipman, 1988), was used to assess similarities of the indicated proteins. Comparison scores in standard deviations, using the RDF2 program (Pearson & Lipman, 1988), and 150 shuffles, are given in square brackets below the values for percent identity.

	XylR(BsuII) (384)	**XylR(Lpe) (388)**	**XylR(Sxy) (383)**	**XylR(Tba) (399)**	**NagC(Eco) (406)**	**ScrK(Smu) (293)**	**FruK(Zmo) (301)**	**Glck(Sco) (317)**	**Orf309(Eco) (309)**	**Orf260(Eco) (260)**	**Orf182(Cpe) (182)**
XylR(BsuI) (384)	93 (384) [145]	30 (388) [34]	31 (383) [39]	39 (395) [65]	26 (394) ([53]	21 (293) [15]	22 (288) [18]	34 (209) [40]	22 (310) [20]	25 (215) [15]	21 (169) [10]
XylR(BsuII) (384)		28 (380) [39]	29 (381) [38]	38 (393) [60]	25 (394) [43]	20 (293) [20]	22 (284) [18]	33 (209) [42]	21 (310) [19]	25 (215) [17]	21 (169) [9]
XylR(Lpe) (388)			37 (374) [54]	31 (240) [29]	21 (382) [22]	24 (34) [−0·5]	16 (209) [7]	27 (161) [13]	19 (64) [0·5]	17 (71) [0·6]	23 (133) [7]
XylR(Sxy) (383)				35 (247) [35]	25 (274) [22]	24 (66) [0·6]	12 (26) [−1]	23 (162) [15]	23 (103) [2]	16 (239) [9]	27 (124) [6]
XylR(Tba) (399)					26 (394) [36]	27 (168) [11]	28 (115) [10]	32 (317) [43]	19 (287) [21]	29 (216) [18]	23 (92) [9]
NagC(Eco) (406)						26 (180) [8]	27 (187) [12]	24 (314) [22]	23 (231) [20]	24 (227) [20]	14 (97) [0·1]
ScrK(Smu) (293)							40 (293) [57]	25 (206) [17]	22 (251) [11]	27 (33) [3]	26 (130) [10]
FruK(Zmo) (301)								28 (189) [17]	21 (259) [6]	22 (125) [6]	26 (174) [15]
GlcK(Sco) (317)									25 (262) [16]	25 (259) [10]	36 (470) [9]
Orf309(Eco) (309)										26 (235) [13]	26 (154) [11]
Orf260(Eco) (260)											20 (94) [8]

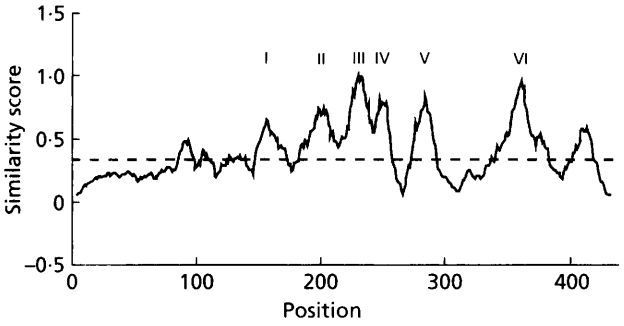

Fig. 1. Average similarity profile of protein members of the ROK family. The average similarity was calculated using a sliding window of 20 amino acids. The average similarity across the entire alignment is shown as a dashed line. Conserved regions of high similarity are indicated (I–VI).

1991; S. Hastrup, GenBank A00033, patent, unpublished; Dwivedi *et al.*, GenBank L18965, unpublished; see also the footnote to Table 1), all from Gram-positive bacteria, and the *N*-acetylglucosamine repressor (NagC) of *Escherichia coli* (Plumbridge, 1989). All of these proteins are about 400 (383–406) amino acids in length. The three homologous sugar kinases include a glucokinase of *Streptomyces coelicolor* (Angell *et al.*, 1992), a fructokinase of *Zymomonas mobilis* (Zembrzuski *et al.*, 1992), and a fructokinase encoded within the sucrose regulon of *Streptococcus mutans* (Sato *et al.*, 1993; Table 1). These three kinases are of about 300 (293–317) residues. Finally, the ROK family includes three ORFs of unknown function. One of these, from the *E. coli* 89–93 min genomic region (Blattner *et al.*, 1993), is of the same size as the sugar kinases. The other two ORFs represent partial sequences. One is encoded downstream of the *araJ* gene of *E. coli*, possibly as part of the *ara* regulon (Reeder & Schleif, 1991), whereas the other is distantly linked to the *nagH* gene coding for a hyaluronidase of *Clostridium perfringens* (Carnard *et al.*, GenBank M81878, unpublished).

Table 2 presents the binary percentage identity and RDF2 comparison scores for most of the sequenced protein members of the ROK family. The high comparison scores obtained when the sequences of the six analysed repressor proteins are compared ($\geqslant$ 22 SD) establish that they are homologous. Similarly, the three kinases [FruK(Zmo), GlcK(Sco) and ScrK(Smu)] are clearly homologous (comparison scores of $\geqslant$ 17 SD). Finally, the three unidentified ORFs, Orf309(Eco), Orf260(Eco) and Orf182(Cpe), are homologous with each other (comparison scores of 8–13 SD). Representative intercluster comparison scores ($>$ 20 SD) establish that these three groups of proteins, the repressors, the kinases and the ORFs, are all homologous.

A multiple alignment of these sequences revealed that the repressors all possess N-terminal extensions of about 80 residues that are lacking in the kinases and the functionally uncharacterized ORFs. An average similarity plot for all members of the family is shown in Fig. 1. The N-terminal 80 residue regions gave low average similarity values since they are not present in the kinases and the functionally uncharacterized ORFs. However, the adjacent regions (alignment positions 140–420 in Fig. 2) show striking sequence similarity. Six peaks of sequence conservation are displayed in Fig. 1 (I–VI). The multiple alignment and the consensus sequence of the twelve proteins is presented in Fig. 2. Three residues at distant positions, the glutamyl residue at alignment position 162, the prolyl residue at position 261, and the glycyl residue at position 319, are fully conserved (see asterisks below the consensus sequence).

The multiple alignment shown in Fig. 2 lacks the N-terminal domains of the repressor proteins. A DNA-binding motif has been noted in this region (Kreuzer *et al.*, 1989; Dodd & Egan, 1990; Sizemore *et al.*, 1991; Lokman *et al.*, 1991; Scheler *et al.*, 1991; Angell *et al.*, 1992). We have analysed all protein members of the ROK family for the presence of a helix-turn-helix (HTH) motif using the method of Dodd & Egan (1990). A recognizable HTH motif ($>$ 2·5 SD) was identified in each of the repressor proteins, in equivalent positions. It is noteworthy that the strongly conserved residues are approximately equidistant throughout the HTH sequence and its flanking regions (see Fig. 3). No recognizable HTH motif was found in the other members of the family or elsewhere within the repressor proteins.

The phylogenetic tree for the 12 protein members of the ROK family is shown in Fig. 4. It can be seen that the six repressors comprise one coherent cluster while the three identified kinases comprise a looser cluster. ORF309(Eco) and the two ORF fragments, Orf260(Eco) and Orf182(Cpe), are most distant from the other members of the family as expected in part from the fact that only partial sequences are available for the latter two proteins.

DISCUSSION

We have characterized a novel family of proteins, which we have designatd the ROK family. It includes transcriptional repressors with N-terminal DNA-binding domains (Fig. 3) and C-terminal sugar-binding domains (Fig. 2), as well as sugar kinases and functionally unidentified ORFs. The kinases are shorter than the repressors by about 80 residues due to the absence of the N-terminal DNA-binding domains. A large and diverse superfamily of proteins including sugar kinases, actin, hsp70 and specific phosphatases has been described in previous publications (Bork *et al.*, 1992, 1993; Reizer *et al.*, 1993). Since the kinases included in the ROK family are likely to belong to this superfamily, our study suggests that transcriptional repressors are also members of this superfamily.

The multiple alignments presented in Figs 2 and 3 as well as the similarity plot (Fig. 1) should provide a guide to structure–function analyses employing techniques such as site-specific mutagenesis, domain fusion of the individual

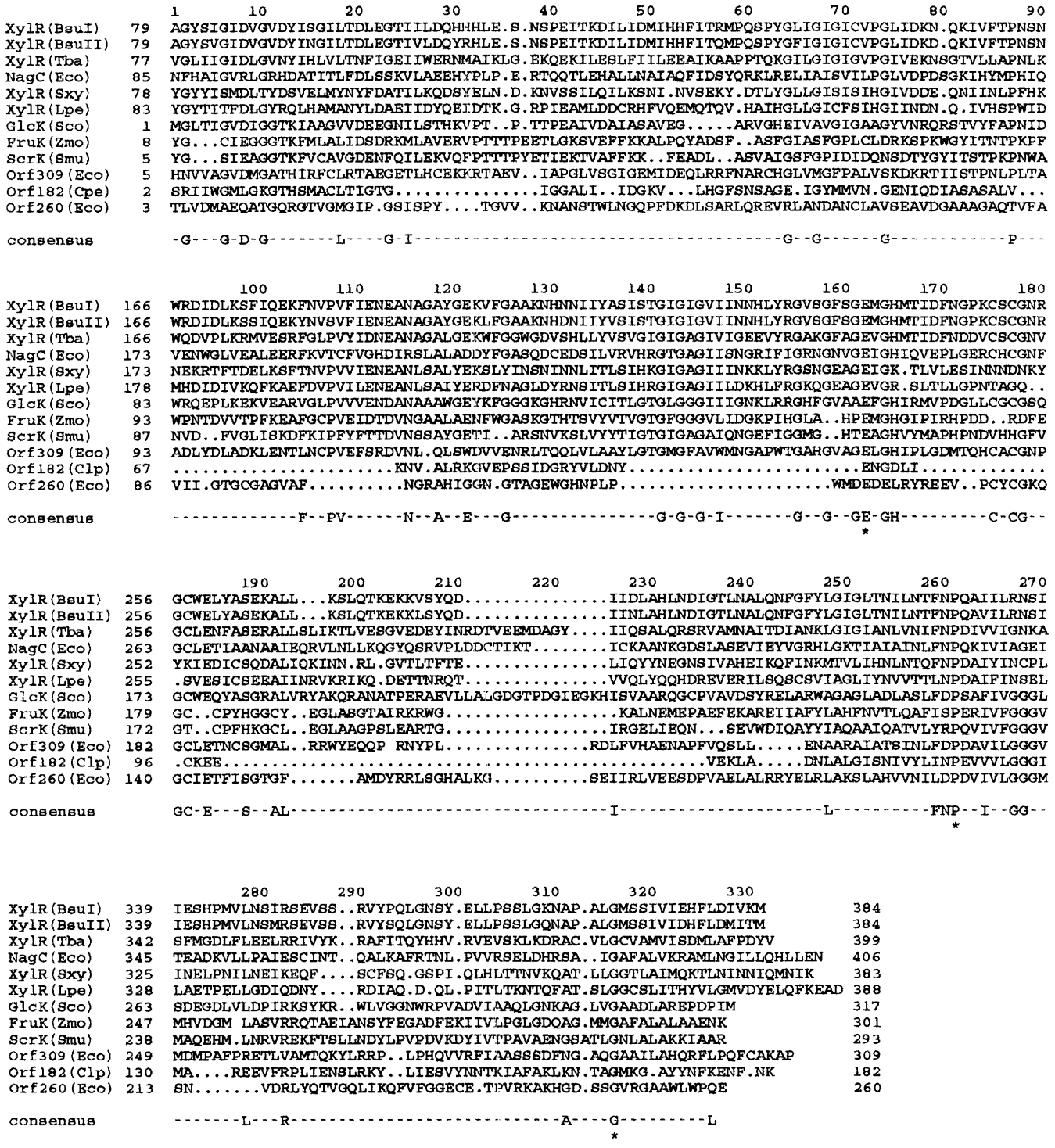

Fig. 2. Multiple alignment of the 12 members of the ROK family. Numbers at the top of the aligned sequences denote the residue position in the multiple alignment. The residue number of each protein is provided at the beginning of each line. A consensus sequence (at least seven residues conserved) is shown below the multiple alignment. Residues conserved in all proteins are marked with asterisks.

proteins and domain swapping between different members of the family. Thus, catalytic residues involved in ATP-binding and phosphoryl transfer might be expected to be conserved only in the kinases, while sugar-binding residues and structural residues involved in maintenance of the three-dimensional domain structures required for sugar binding are more likely to be conserved in the repressors as well as the kinases. The fact that the *E. coli*

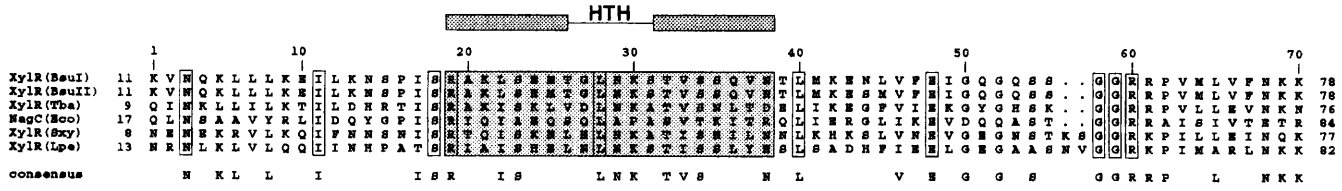

Fig. 3. Multiple alignment of the DNA-binding domains of six repressor proteins. Numbers at the top of the aligned sequences denote the residue position in the multiple alignment. The residue numbers of each protein are provided at the beginning and at the end of each line. Identical residues are boxed, while a consensus sequence (at least four residues conserved) is provided below the multiple alignment. The region containing the helix-turn-helix (HTH) motif is stippled.

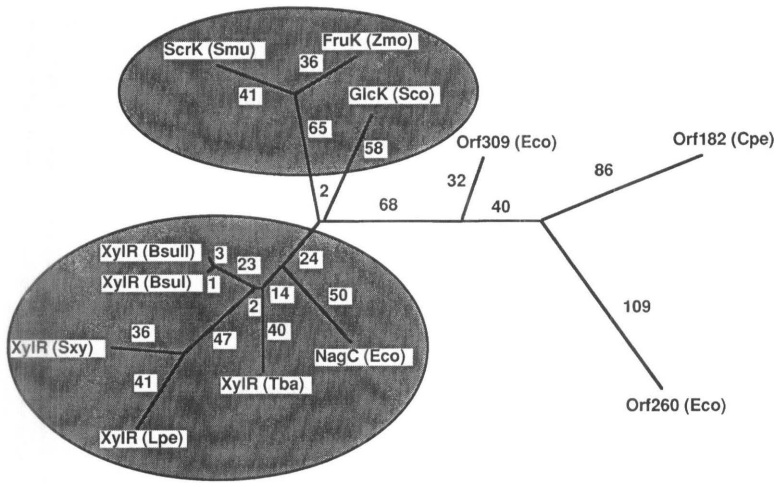

Fig. 4. Phylogenetic tree of the 12 proteins that comprise the ROK family. The programs of Feng & Doolittle (1990) were used for calculation of the branch lengths and construction of the phylogenetic tree. Relative evolutionary distances are given adjacent to the branches. Abbreviations are as listed in the legend to Table 1. The repressor and sugar kinase clusters are highlighted by a stippled area.

Orf309 lacks the N-terminal DNA-binding domain suggests that it is a sugar kinase rather than a repressor protein. The finding that Orf309(Eco) is encoded in an operon with genes encoding proteins homologous to pentose epimerase as well as a sugar transport protein (unpublished results) supports this suggestion. These predictions should serve as a guide to future biochemical analyses aimed at identifying the functions of these putative proteins.

The ROK family includes the third recognized group of homologous repressor proteins which possesses ligand-binding domains that share a common origin with proteins of unrelated function. It is the first family whose members exhibit catalytic rather than receptor activity (Tam & Saier, 1993a). It is relevant to note, however, that one of the bacterial periplasmic receptor families includes a catalytic protein (Tam & Saier, 1993b). Our observation that transcriptional regulatory proteins and enzymes share a common origin is in accordance with an early prediction (Saier & Jacobson, 1984) as well as current concepts of the modular origin of proteins (Doolittle & Bork, 1993; Saier, 1994).

ACKNOWLEDGEMENTS

We thank Mary Beth Hiller for expert assistance in the preparation of this manuscript. This work was supported by Public Health Service grants 5RO1 AI 21702 and 2RO1 AI 14176 from the National Institute of Allergy and Infectious Diseases. Friedrich Titgemeyer was supported by a fellowship from the Alexander von Humboldt Foundation of Germany.

REFERENCES

Angell, S., Schwarz, E. & Bibb, M. J. (1992). The glucose kinase gene of *Streptomyces coelicolor* A3(2): its nucleotide sequence, transcriptional analysis and role in glucose repression. *Mol Microbiol* **6**, 2833–2844.

Blattner, F. R., Burland, V. D., Plunkett, G., III, Sofia, H. J. & Daniels, D. L. (1993). Analysis of the *Escherichia coli* genome. IV. DNA sequence of the region from 89·2 to 92·8 minutes. *Nucleic Acids Res* **21**, 5408–5417.

Bork, P., Sander, C. & Valencia, A. (1992). An ATPase domain common to prokaryotic cell cycle proteins, sugar kinases, actin, and hsp70 heat shock proteins. *Proc Natl Acad Sci USA* **89**, 7290–7294.

Bork, P., Sander, C. & Valencia, A. (1993). Convergent evolution of similar enzymatic function on different protein folds: the hexokinase, ribokinase, and galactokinase families of sugar kinases. *Prot Sci* **2**, 31–40.

Dodd, I. B. & Egan, J. B. (1990). Improved detection of helix-turn-helix DNA-binding motifs in protein sequences. *Nucleic Acids Res* **18**, 5019–5026.

Doolittle, R. F. & Bork, P. (Oct. 1993). Evolutionarily mobile modules in proteins. *Sci Am*, 50–56.

Feng, D.-F. & Doolittle, R. F. (1990). Progressive alignment and phylogenetic tree construction of protein sequences. *Methods Enzymol* **183**, 375–387.

Kreuzer, P., Gaertner, D., Allmansberger, R. & Hillen, W. (1989).

Identification and sequence analysis of the *Bacillus subtilis* W23 *xylR* gene and *xyl* operator. *J Bacteriol* **171**, 3840–3845.

Lokman, B. C., van Santen, P., Verdoes, J., Kruese, J., Leer, R. J., Posno, M. & Pouwels, P. H. (1991). Organization and characterization of three genes involved in D-xylose catabolism in *Lactobacillus pentosus*. *Mol & Gen Genet* **230**, 161–169.

Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* **85**, 2444–2448.

Plumbridge, J. (1989). Sequence of the *nag BACD* operon in *E. coli* K12 and pattern of transcription within the *nag* regulon. *Mol Microbiol* **3**, 505–515.

Reeder, T. & Schleif, R. (1991). Mapping, sequence, and apparent lack of function of *araJ*, a gene of the *Escherichia coli* arabinose regulon. *J Bacteriol* **173**, 7765–7771.

Reizer, J., Reizer, A. & Saier, M. H., Jr (1993). Exopolyphosphate phosphatase and guanosine pentaphosphate phosphatase belong to the sugar kinase/actin/hsp70 superfamily. *TIBS* **18**, 247–248.

Rygus, T., Scheler, A., Allmansberger, R. & Hillen, W. (1991). Molecular cloning, structure, promoters and regulatory elements for transcription of the *Bacillus megaterium* encoded regulon for xylose utilization. *Arch Microbiol* **155**, 535–542.

Saier, M. H., Jr (1994). Computer-aided analyses of transport protein sequences: gleaning evidence concerning function, structure, biogenesis, and evolution. *Microbiol Rev* **58**, 71–93.

Saier, M. H., Jr & Jacobson, G. R. (1984). *The Molecular Basis of Sex and Differentiation: a Comparative Study of Evolution, Mechanism and Control in Microorganisms*. New York: Springer-Verlag.

Sato, Y., Yamamoto, Y., Kizaki, H. & Kuramitsu, H. K. (1993). Isolation, characterization and sequence analysis of the *scrK* gene encoding fructokinase of *Streptococcus mutans*. *J Gen Microbiol* **139**, 921–927.

Scheler, A., Rygus, T., Allmansberger, R. & Hillen, W. (1991). Molecular cloning, structure, promoters and regulatory elements for transcription of the *Bacillus licheniformis* encoded regulon for xylose utilization. *Arch Microbiol* **155**, 526–534.

Sizemore, C., Buchner, E., Rygus, T., Witke, C., Gotz, F. & Hillen, W. (1991). Organization, promoter analysis and transcriptional regulation of the *Staphylococcus xylosus* xylose utilization operon. *Mol & Gen Genet* **227**, 377–384.

Tam, R. & Saier, M. H., Jr (1993a). Structural, functional, and evolutionary relationships among extracellular solute-binding receptors of bacteria. *Microbiol Rev* **57**, 320–346.

Tam, R. & Saier, M. H., Jr (1993b). A bacterial periplasmic receptor homologue with catalytic activity: cyclohexadienyl dehydratase of *Pseudomonas aeruginosa* is homologous to receptors specific for polar amino acids. *Res Microbiol* **144**, 165–169.

Vartak, N. B., Reizer, J., Reizer, A., Gripp, J. T., Groisman, E. A., Wu, L.-F., Tomich, J. M. & Saier, M. H., Jr (1991). Sequence and evolution of the FruR protein of *Salmonella typhimurium*: a pleiotropic transcriptional regulatory protein possessing both activator and repressor functions which is homologous to the periplasmic ribose-binding protein. *Res Microbiol* **142**, 951–963.

Weickert, M. J. & Adhya, S. (1992). A family of bacterial regulators homologous to Gal and Lac repressors. *J Biol Chem* **267**, 15869–15874.

Wilson, S. & Drew, R. (1991). Cloning and DNA sequence of *amiC*, a new gene regulating expression of the *Pseudomonas aeruginosa* aliphatic amidase, and purification of the *amiC* product. *J Bacteriol* **173**, 4914–4921.

Zembrzuski, B., Chilco, P., Liu, X.-L., Liu, J., Conway, T. & Scopes, R. K. (1992). The fructokinase gene from *Zymomonas mobilis*: cloning, sequencing, expression and structural comparison of the enzyme with other hexose kinases. *J Bacteriol* **174**, 3455–3460.

Received 7 February 1994; accepted 29 April 1994.