Novel phosphotransferase genes revealed by bacterial genome sequencing: a gene cluster encoding a putative *N*-acetylgalactosamine metabolic pathway in *Escherichia coli*

Jonathan Reizer, Tom M. Ramseier, Aiala Reizer, Alain Charbit 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 analysed a gene cluster in the 67.4–76.0 min region of the *Escherichia coli* chromosome, revealed by recent systematic genome sequencing. The genes within this cluster include: (1) five genes encoding homologues of the *E. coli* mannose permease of the phosphotransferase system (IIB, IIB', IIC, IIC' and IID); (2) genes encoding a putative *N*-acetylgalactosamine 6-phosphate metabolic pathway including (a) a deacetylase, (b) an isomerizing deaminase, (c) a putative carbohydrate kinase, and (d) an aldolase; and (3) a transcriptional regulatory protein homologous to members of the DeoR family. Evidence is presented suggesting that the aldolase-encoding gene within this cluster is the previously designated *kba* gene that encodes tagatose-1,6-bisphosphate aldolase. These proteins and a novel IIA^{Man}-like protein encoded in the 2.4–4.1 min region are characterized with respect to their sequence similarities and phylogenetic relationships with other homologous proteins. A pathway for the metabolism of *N*-acetylgalactosamine biochemically similar to that for the metabolism of *N*-acetylglucosamine is proposed.

Keywords: *Escherichia coli* genome, phosphotransferase system, sugar transport, *N*-acetylgalactosamine, aldolases

INTRODUCTION

Systematic sequencing of the *Escherichia coli* genome has revealed a number of previously unrecognized genes encoding proteins of the phosphoenolpyruvate:sugar phosphotransferase system (PTS) (Reizer *et al.*, 1994a, b, 1995; unpublished observations). These genes are generally found within operons concerned with carbohydrate metabolism. *In silico* analyses have revealed the probable functions of many of these PTS proteins as well as of putative catabolic enzymes encoded by genes within the corresponding operons. Genes encoding PTS proteins are scattered randomly on the circular chromosome and represent about 1% of all currently sequenced *E. coli* genes. Many more such genes will probably be found as sequencing of the *E. coli* genome is completed.

Abbreviations: CAI, codon adaptation index; FBP, fructose 1,6-bisphosphate; MDA, metal-dependent (class II) aldolase; PEP, phosphoenolpyruvate; PTS, phosphoenolpyruvate:sugar phosphotransferase system.

The sugar-specific PTS proteins which comprise the Enzyme II complexes consist of three (or four) proteins or protein domains termed IIA, IIB, IIC (and sometimes IID) (Saier & Reizer, 1992, 1994; Postma *et al.*, 1993). IIA and IIB are phosphoryl transfer proteins or protein domains that are either cytoplasmic or localized to the cytoplasmic side of the bacterial membrane. The IIC and IID proteins/domains are integral membrane constituents that span the membrane. Recent sequence and 3-dimensional structural analyses have established that some IIA proteins and some IIB proteins are non-homologous and possess completely different secondary structures (reviewed in Saier & Reizer, 1994). This observation has led to the conclusion that the PTS is a mosaic system derived from several different sources. PTS protein domains, homologous to domains which function as phosphoryl transfer constituents of Enzyme II complexes, have been identified in non-PTS permeases and catabolic enzymes where they are thought to serve regulatory functions (Saier & Reizer, 1994).

Early sequencing analyses revealed that the mannose Enzyme II complex of *E. coli*, of broad sugar-specificity (Rephaeli & Saier, 1980), consisted of four domains IIA, IIB, IIC and IID that were not homologous to any of the previously sequenced PTS permeases (Erni *et al.*, 1987). Subsequent sequence analyses revealed that the fructose Enzyme II complex of *Bacillus subtilis* (Martin-Verstraete *et al.*, 1990) and the sorbose Enzyme II complex of *Klebsiella pneumoniae* (Wehmeier & Lengeler, 1994; Wehmeier *et al.*, 1995) were homologous to the *E. coli* mannose system (Reizer *et al.*, 1991b; Lengeler *et al.*, 1994). These three Enzyme II complexes comprised the so-called 'splinter group' of PTS permeases that differ in structure and phylogenetic origin from other sequenced PTS permeases (Erni, 1992; Saier & Reizer, 1992). It has been suggested that the mechanism of action of the 'splinter group' permeases (which all possess IIC and IID transmembrane proteins) may differ from that of the other permeases which lack a IID protein. Moreover, the two phosphorylation sites of the mannose permease are two histidine residues, whereas a histidine and a cysteine are phosphorylated in other PTS permeases (Erni *et al.*, 1989; Pas *et al.*, 1991; Meins *et al.*, 1993; Stolz *et al.*, 1993).

Recent sequence data from F. R. Blattner and co-workers in the 67.4–76.0 min region of the *E. coli* genome (GenBank accession no. U18997) has revealed a cluster of genes encoding at least five novel PTS proteins. All of the encoded PTS proteins exhibit striking sequence similarities with proteins of the 'splinter group' of PTS permease proteins. Analyses of additional genes in this cluster have revealed the presence of catabolic enzymes which could (with one exception) be assigned to a single coherent metabolic pathway which we predict is concerned with *N*-acetylgalactosamine metabolism. We therefore suggest that this gene cluster, which we here designate the *aga* (*N*-acetylgalactosamine) gene cluster, is potentially concerned with the metabolism of this sugar. A detailed description of this gene cluster and of the encoded protein products are reported here.

METHODS

Computer-aided analyses. All database searches and sequence analyses were performed using the BLAST program (Altschul *et al.*, 1990), the GCG package from the University of Wisconsin (Devereux *et al.*, 1984) and the DNASYSTEM package (Smith, 1988). Comparison scores (expressed in SD) were calculated using the RDF2 program with 150 shuffles as indicated in the table legends. A value of 6 SD is suggestive of homology whereas a value of 9 SD establishes homology. Construction of phylogenetic trees and estimation of the relative evolutionary distances among members of a protein family were as described by Reizer & Reizer (1994) using the progressive alignment method of Feng & Doolittle (1990). Mean hydropathy and mean similarity were calculated using a sliding window of 20 residues. Hydropathy plots were designed according to the method of Kyte & Doolittle (1982). The calculated mean similarity at a position in a multiple alignment was an arithmetic mean of all pairwise comparisons. The mean similarity across the entire multiple alignment was the sum of similarities obtained in all windows divided by the number of windows.

RESULTS

The *aga* gene cluster

Fig. 1 shows the gene order, orientation and structure of the *aga* gene cluster of *E. coli*, recently sequenced in the 67.4–76.0 min chromosomal region as part of the *E. coli* genome sequencing project (GenBank accession no. U18997; submitted by G. Plunkett). The figure presents (1) a restriction map of the gene region (top), (2) the proposed gene designations (in bold italics), (3) the original gene designations (in parentheses), (4) the functional assignments of the protein products when known, based on homology to previously characterized proteins, (5) the codon adaptation indices [CAI (Sharp & Li, 1987)] of the ORFs, (6) the G + C content (G + C mol %) of each gene; (7) information regarding the intergenic regions (in brackets), and (8) putative RBSs of the genes included in the *aga* locus (see legend to Fig. 1). This cluster contains 11 clearly identifiable ORFs as follows (predicted functions of protein products are given in parentheses): (1) *agaR* (a regulatory protein); (2) *agaZ* (a putative sugar kinase homologous to GatZ); (3) *agaB* and *agaV* (two IIB^{Man} homologues); (4) *agaC* and *agaW* (two IIC^{Man} homologues); (5) *agaD* (IID^{Man} homologue); (6) *agaA* (putative deacetylase); (7) *agaI* (putative isomerizing deaminase); (8) *agaY* (putative aldolase); and (9) *agaS* (protein homologous to the ketose–aldose isomerase-bearing domain of characterized glutamine:fructose amidotransferases). CAI values suggest that these genes are expressed at low-to-moderate levels with substantial variation within the gene cluster. The G + C content is, in general, somewhat higher than that for the *E. coli* genome although the G + C content of one gene (*agaB*) is substantially lower (44%) than that for the *E. coli* genome.

PTS Enzyme II complexes (PTS permeases) include IIA, IIB, IIC (and sometimes IID) proteins or domains usually encoded within a single operon or gene cluster. However, in a few cases, such as the *glc* operon (*ptsG*) of *E. coli*, the twin *sac* genes (*sacP* and *sacX*) of *B. subtilis*, the *scr* operon (*scrA*) of enteric bacteria, and the *tre* operon (*treB*) of *E. coli*, the IIA encoding genes do not cluster with those encoding the other permease domains. In the *aga* gene cluster, no IIA homologue is found. However, a putative IIA^{Aga} is found in the 2.4–4.1 min region of the *E. coli* genome (Fujita *et al.*, 1994), distant from the *aga* gene cluster. Below we report the results of our analyses of these ORFs, presenting the analyses of first, the *AgaR* regulatory protein, second, the PTS proteins, and third, the putative catabolic enzymes.

AgaR (ORF f269)

The *agaR* gene, divergently transcribed from the other genes in the *aga* gene cluster (Fig. 1), encodes a homologue of a family of transcriptional regulatory proteins known as the DeoR family. All characterized protein members of this family control sugar-catabolic operons which appear to be induced by a phosphorylated sugar generated by

Fig. 1. Restriction and genetic map of the *aga* gene cluster. The genetic map shows the proposed gene assignments, the number of amino acids in each ORF (in parentheses), the transcriptional orientation (designated 'o' or 'f') of these ORFs, the designations of the encoded proteins, and the numbers of base pairs (in brackets) in the intergenic regions (a minus sign denotes an overlapping region). The calculated CAI (Sharp & Li, 1987) and the G+C content of each gene, are indicated at the bottom of the corresponding map. The dashed regions in ORF o157 and ORF o263 denote the previously included 5'-regions in these ORFs which had been termed ORF o169 and ORF o290, respectively (see text for details). The putative RBSs of the 11 genes comprising this cluster are indicated by roman numerals as follows: I, AAGGA; II, GGAG; III, GAGG; IV, GAGG; V, GGTG; VI, TAAGGA; VII, GAGG; VIII, GAGG; IX, GAGG; X, GAGG; XI, not detected. Restriction site abbreviations are: B, *Bam*HI; Bg, *Bgl*I; E, *Eco*RI; EV, *Eco*RV; K, *Kpn*I; P, *Pst*I; Pc, *Pac*I; Pm, *Pme*I; Pv, *Pvu*I.

enzymes comprising metabolic pathways encoded within these operons. As summarized in Table 1, AgaR is most similar to GutR, the repressor of the *E. coli* glucitol operon (35% identity; 51 sd; Yamada & Saier, 1988). Recent studies with the LacR repressor of *Lactococcus lactis* have suggested that the lysine residues at positions 72, 80 and 213, as well as aspartate₂₁₀ function in binding of the inducer tagatose 6-phosphate (van Rooijen, 1993). Thus, the mutant proteins K72A, K80A, K213A and D210A strongly depressed expression of the *lac* operon in cultures grown on lactose, and addition of tagatose 6-phosphate failed to dissociate the complexes formed between any of these mutant repressors and DNA fragments bearing the *lac* promoter/operator region. Since lysine₇₂ is conserved in AgaR, and lysine₈₀, lysine₂₁₃ and aspartate₂₁₀ are fully conserved in all protein members of this family, it is likely that these residues form part of a common inducer-binding site in LacR, AgaR and the other 12 sequenced repressors that comprise this family.

A signature sequence of the DeoR family has been derived (unpublished observations). This modified signature sequence which contains the helix-turn-helix of these transcriptional regulators is RX₃[LIVM]X₃[LIVM]X₂₀T[LIVMA]R[RKAN]D[LIVMF]. The amino acid at a position in which the residue is not specified is denoted by X, whereas residues in brackets indicate alternative possibilities at a single position.

The alignment of the N-terminal helix-turn-helix DNA-binding regions of proteins comprising the DeoR family is shown in Fig. 2. The alignment shows that only the

second half of this motif is well conserved in all of the AgaR-related repressors.

While this work was in progress, an additional protein member of the DeoR family was identified in *B. subtilis*. Interestingly, the *B. subtilis* protein was shown to function together with LevR in transcriptional regulation of the *lev* operon which encodes fructose-specific PTS proteins homologous to the Aga-specific PTS constituents described below (P. Glaser, I. Stulke & I. Martin-Verstraete, 8th International Conference on Bacilli, Stanford, CA, USA, July 1995).

AgaB and AgaV (ORF o158 and ORF o157, respectively)

Two genes in the *aga* gene cluster, *agaB* and *agaV*, encode IIB^{Man} homologues, IIB^{Aga} and IIB^{Aga'}, respectively, and as revealed in Table 2, they are more similar to each other (46% identity; 44 sd) than to other members of the family. The two IIB^{Aga} proteins are clearly homologous to the IIB domains of (a) the *E. coli* mannose PTS [IIB^{Man}(Eco); 34–36% identity, 27–29 sd], (b) the *K. pneumoniae* sorbose PTS [IIB^{Sor}(Kpn); 30–35% identity, 28–31 sd] and (c) the *B. subtilis* fructose PTS [IIB^{Fru}(Bsu); 27–32% identity, 23–27 sd] (Table 2). It should be noted that IIB^{Aga'} resembles the three homologues more than does IIB^{Aga}, suggesting that it may play a primary role as a phosphoryl transfer protein (see Discussion).

AgaV was originally designated ORF o169 (GenBank accession no. U18997). We propose that the translational

Table 1. Binary comparison of AgaR (ORF f269) with protein members of the DeoR family

The FASTA program (Pearson & Lipman, 1988), was used to assess similarities for segments having the indicated number of compared residues. Comparison scores in standard deviations were determined using the RDF2 program (Pearson & Lipman, 1988) and 150 shuffles of the segments compared. Abbreviations and references to the published sequences are as follows: GutR(Eco), repressor of the glucitol operon of *E. coli* (Yamada & Saier, 1988); AccR(Atu), *Agrobacterium tumefaciens* repressor of opine catabolism and nopaline Ti plasmid conjugal transfer (von Bodman *et al.*, 1992); GatR(Eco), transcriptional regulator of galactitol utilization of *E. coli* (Komoda *et al.*, 1991; Nobelmann & Lengeler, 1995); GlpR(Eco), repressor of the glycerol 3-phosphate regulon of *E. coli* (Choi *et al.*, 1988); ORF f251(Eco) and ORF o269(Eco), two putative transcriptional regulators that are encoded respectively within the 92·8–100 min, GenBank accession no. U14003 (unpublished results cited in Burland *et al.*, 1995), and the 87·2–89·2 min (Plunkett *et al.*, 1993) regions of the *E. coli* chromosome; LacR(Lla), -(Sau) and -(Smu), repressor of the lactose operons of *L. lactis* (van Rooijen *et al.*, 1991), *Staph. aureus* (Oskouian & Stewart, 1990), and *Strep. mutans* (Rosey & Stewart, 1992), respectively; DeoR(Eco), repressor of *deoCABD* that encodes nucleoside-catabolizing enzymes in *E. coli* (Valentin-Hansen *et al.*, 1985); FucR(Eco), transcriptional regulator (activator) of L-fucose metabolism in *E. coli* (Lu & Lin, 1989); SgcR(Eco), putative transcriptional regulator in the *sgc* gene cluster of *E. coli* (GenBank accession no. U14003; Burland *et al.*, 1995; unpublished results); ORF232(Mca), hypothetical protein of *Myc. capricolum* (Bork *et al.*, 1995).

Protein	Accession no.	Identity (%)	Segments compared (no. of residues)	Comparison score (SD)
GutR(Eco)	P15082	35	254	51
AccR(Ata)	P32104	30	239	37
GatR(Eco)	P36930	39	183	36
GlpR(Eco)	P09392	28	251	34
ORF o269(Eco)	P32144	27	252	29
ORF232(Mca)	S48604	26	227	28
LacR(Lla)	P18816	28	244	28
DeoR(Eco)	P06217	27	230	28
SgcR(Eco)	P39361	26	251	27
FucR(Eco)	P11554	26	239	24
LacR(Sau)	P16644	25	232	23
LacR(Smu)	P26422	26	245	20
ORF f251(Eco)	P39299	27	253	22

Fig. 2. Helix–turn–helix motifs of the 14 homologous DNA-binding proteins of the DeoR family. The consensus sequence is shown below the aligned sequences, and residues conserved in all 14 proteins are highlighted (black background). The residue number is shown at the beginning of the sequence for each protein whereas numbers at the top of the aligned sequences denote the residue position in the multiple alignment. Abbreviations and references to the published sequences are as described in the legend to Table 1.

initiation codon (GTG) assignment of ORF o169 is incorrect and that the actual gene should be designated *orf* o157 beginning with an ATG codon, 36 bp downstream from the previously proposed (GTG) initiation codon. Our proposal is based on the following considerations: (1) a recognizable Shine–Dalgarno sequence is absent in the 5′-region preceding the previously designated translational start GTG codon, whereas the initiation codon (ATG) proposed here is preceded (8 bp spacing) by a conserved Shine–Dalgarno sequence, i.e. GAGG; (2) ATG is used in preference to GTG in *E. coli*; (3) the previously proposed gene encoding ORF o169 overlaps at its 5′-end with the 3′-end of the gene encoding ORF o426 (14 bp overlap) whereas our proposed assignment (ORF o157) of the former ORF eliminates this overlap; and (4) none of the IIB homologues exhibits an N-terminal sequence corresponding to the one found in the putative ORF o169.

The multiple alignment of the five members of this family

Table 2. Comparison of the amino acid sequences of the duplicated IIB^{Aga} enzymes with the mannose-, sorbose- and fructose-specific Enzymes IIB

The FASTA program (Pearson & Lipman, 1988) using the dipeptide identities mode was used to assess similarities for segments having the indicated number of compared residues. Comparison scores in standard deviations were determined using the RDF2 program (Pearson & Lipman, 1988) and 150 shuffles of the segments compared. Abbreviations and references to the published sequences are as follows: IIB^{Man}, mannose-specific IIB protein domain of *E. coli* [IIB^{Man}(Eco), i.e. the C-terminal portion, previously designated the P20 protein domain of Enzyme IIB^{Man}; Erni *et al.*, 1987, 1989]; IIB^{Sor}, sorbose-specific Enzyme IIB of *K. pneumoniae* [IIB^{Sor}(Kpn), Wehmeier & Lengeler, 1994; Wehmeier *et al.*, 1995]; IIB^{Fru}, fructose-specific Enzyme IIB of *B. subtilis* [IIB^{Fru}(Bsu), Martin-Verstraete *et al.*, 1990].

Enzyme IIB	Identity (%)		Segments compared (no. of residues)		Comparison score (SD)	
	IIB ^{Aga'} (157)	IIB ^{Aga} (158)	IIB ^{Aga'} (157)	IIB ^{Aga} (158)	IIB ^{Aga'} (157)	IIB ^{Aga} (158)
IIB ^{Aga'} (157)	NA	46	NA	153	NA	44
IIB ^{Aga} (158)	46	NA	153	NA	44	NA
IIB ^{Man} (Eco) (168)	36	34	155	154	29	27
IIB ^{Sor} (Kpn) (164)	35	30	153	155	31	28
IIB ^{Fru} (Bsu) (163)	32	27	158	156	27	23

NA, Not applicable.

is presented in Fig. 3(a), and the mean similarity plot for these proteins is shown in Fig. 3(b). All five proteins exhibit greatest sequence similarity near their N-termini immediately surrounding the active site, the phosphorylatable histidyl residue of IIB^{Man} (P20) (Erni *et al.*, 1989; Stolz *et al.*, 1993). This region was used to construct a signature sequence for this family of Enzymes IIB (see Fig. 3c). This sequence, R[LIVM]DXR[LIVMF]₂HGQ[LIVM]X₃W, proved to be specific to the five members of the family when Swiss-Prot database (version 31.0) was screened.

AgaC and AgaW (ORF o267 and ORF o133, respectively)

agaC and *agaW* encode two hydrophobic proteins homologous to the sequenced protein members of the Enzyme IIC^{Man} family. The protein product of *agaC* (IIC^{Aga}) is a full-length protein similar in length to the mannose, sorbose and fructose specific IIC homologues (266–269 residues; see Table 3). By contrast, IIC^{Aga'}, the protein product of *agaW*, is a truncated IIC protein of only 133 amino acids, corresponding to the N-termini of its four full-length IIC homologues. The comparison scores recorded in Table 3 (12–26 SD) establish that IIC^{Aga'} is most similar to IIC^{Aga}, and that all five proteins are homologous.

Fig. 4(a) shows the multiple alignment of the five homologous members of the IIC^{Man} family. The most conserved portion of the alignment is the N-terminal region preceding the conserved glutamate at position 58 in the multiple alignment shown in Fig. 4(a). The sequence including this glutamate, GGTLE, is fully conserved except in IIC^{Aga} in which the sequence is GGLTE. These sequences may correspond to the well-conserved motif found in most of the other IIC proteins of the PTS (Reizer *et al.*, 1991b). The glutamyl residue in the GGIHE sequence of the mannitol IIC protein was recently shown to be essential for sugar binding and catalytic function (Jacobson & Saraceni-Richards, 1993). We note, however, that the region we propose for this essential glutamate in the IIC^{Man} family of proteins differs from the recently proposed sequence KLTEG (Lengeler *et al.*, 1994; Wehmeier *et al.*, 1995), which is conserved in the mannose-, fructose-, and sorbose-specific IID proteins but not in the homologous IID^{Aga} (see below).

Within the N-terminal regions of the IIC^{Man} family proteins is a cysteinyl residue (alignment position 36) which is conserved in IIC^{Man}, IIC^{Fru}, IIC^{Sor}, and IIC^{Aga}, but not in the shorter but otherwise homologous IIC^{Aga'} (Fig. 4a). Recent data have demonstrated that substitution of this cysteinyl residue in IIC^{Man} (C36S) does not affect phosphotransferase activity indicating that it is not an essential catalytic residue (Rhie *et al.*, 1994).

Fig. 3. Multiple alignment (a) and mean similarity plot (b) of the proteins comprising the IIB^{Man} family (see Table 2). Fully conserved residues in (a) are boxed; the presumed active site histidyl residue (site of phosphorylation; Erni *et al.*, 1989; Stolz *et al.*, 1993) is highlighted (black background), and residues common in at least three of the five sequenced proteins are presented in the consensus sequence at the bottom. Numbers to the left of the sequences shown denote the residue numbers in the individual proteins. Numbers above the aligned sequences refer to the alignment position and not to any one of the aligned proteins. For references to the published sequences see the legend to Table 2. A sliding window of 20 residues was used to calculate the mean similarity (b) at a particular position. The mean similarity score across the entire alignment is shown by the dashed line. The proposed signature sequence of this IIB^{Man} protein family is shown above the mean similarity plot. Residues in brackets indicate alternative possibilities at a single position, whereas X corresponds to any residue.

Fig. 4(b, c) presents the mean hydropathy and mean similarity plots, respectively, for the four full-length proteins of the IIC^{Man} family. The signature sequence for this family of proteins is shown above the mean similarity plot. This sequence contains the conserved glutamate noted above. The mean hydropathy plot (Fig. 4b) suggests the presence of seven putative transmembrane helical segments and a strongly hydrophilic loop separating the

four N-terminal α -helices from the three C-terminal spanners. Similar results were obtained for the individual protein members of this family when analysed with the TopPred II program (Claros & von Heijne, 1994) which considers the prevalence of positively charged residues in the interior (cytoplasmic) loops (von Heijne, 1994). Use of this 'positive inside rule' further suggested that in all IIC^{Man}-like proteins the N-terminal sequence is localized

Table 3. Comparison of the amino acid sequences of the two IIC^{Aga} proteins with the mannose-, sorbose- and fructose-specific Enzyme IIC proteins

The FASTA program using the dipeptide identities mode (Pearson & Lipman, 1988) was used to assess similarities for segments having the indicated number of compared residues. Comparison scores in standard deviations were determined using the RDF2 program (Pearson & Lipman, 1988) and 150 shuffles of the segments compared. Abbreviations and references to the published sequences are as follows: IIC^{Man}, mannose-specific IIC protein of *E. coli* [IIC^{Man}(Eco), previously named II-P^{Man}, Erni *et al.*, 1987]; IIC^{Sor}, sorbose-specific IIC protein of *K. pneumoniae* [IIC^{Sor}(Kpn), Wehmeier & Lengeler, 1994; Wehmeier *et al.*, 1995]; IIC^{Fru}, fructose-specific IIC protein of *B. subtilis* [IIC^{Fru}(Bsu), Martin-Verstraete *et al.*, 1990].

Enzyme IIC	Identity (%)		Segments compared (no. of residues)		Comparison score (SD)	
	IIC ^{Aga} (133)	IIC ^{Aga} (267)	IIC ^{Aga} (133)	IIC ^{Aga} (267)	IIC ^{Aga} (133)	IIC ^{Aga} (267)
IIC ^{Aga} (133)	NA	39	NA	124	NA	22
IIC ^{Aga} (267)	39	NA	124	NA	22	NA
IIC ^{Man} (Eco) (266)	36	26	108	263	12	26
IIC ^{Sor} (Kpn) (266)	36	27	112	256	14	23
IIC ^{Fru} (Bsu) (269)	38	24	123	261	12	22

NA, Not applicable.

to the cytoplasmic side of the membrane (data not shown). We note that the prediction of seven transmembrane segments in the IIC^{Man}-like proteins, which is based on analyses of the multiple aligned sequences as well as of the individual protein members of this family, is at odds with previous predictions (Lengeler *et al.*, 1994). In accordance with our analyses and predictions, the IIC^{Aga} protein is truncated in the intracellular hydrophilic loop region which separates the two clusters of transmembrane helical segments.

AgaD (ORF o263)

AgaD (IIC^{Aga}) was originally designated ORF o290 (GenBank accession no. U18997). We propose that the actual gene should be designated *orf o263* beginning with an ATG codon, 79 bp downstream from the previously proposed (ATG) initiation codon. Our proposal is based on the following considerations: (a) a recognizable Shine–Dalgarno sequence is absent in the 5'-region preceding the previously designated translational start codon, whereas the initiation codon proposed here is preceded (8 bp spacing) by a conserved Shine–Dalgarno sequence, i.e. GAGG; (b) none of the IIC^{Aga} homologues exhibit an N-terminal extension (27 residues) corresponding to the one found in the previously designated ORF o290 (see Fig. 5a); (c) the CAI value is substantially lower for the previously proposed N-terminal region of ORF

o290 than for ORF o263 (0.16 versus 0.31); (d) the previously proposed gene encoding ORF o290 overlaps at its 5'-end with the 3'-end of the gene encoding ORF o267 (92 bp overlap). This degree of overlap is rarely observed, and our proposed assignment of AgaD (ORF o263) reduces this overlap to 11 bp.

IIC^{Aga} is 34–37% identical throughout its length to the IIC^{Man} of *E. coli* [IIC^{Man}(Eco)], IIC^{Fru} of *B. subtilis* [IIC^{Fru}(Bsu)], and IIC^{Sor} of *K. pneumoniae* [IIC^{Sor}(Kpn)]. The comparison scores obtained (35–38 SD) establish that IIC^{Aga} is homologous to the other three IIC proteins of this family.

Fig. 5 presents the multiple alignment (a), the mean hydropathy plot (b) and the mean similarity plot (c) of this IIC family. Approximately 20% of the residues are fully conserved in the four members of the IIC^{Man} family. The higher similarity among proteins of the IIC^{Man} or IIC^{Man} families, as compared to the similarity among members of the IIC^{Man} family is documented by the mean similarity values calculated for the entire alignment of the IIC^{Man} homologues – 0.85 (dashed line in Fig. 5c), the IIC^{Man} homologues – 0.83 (dashed line in Fig. 4c), and the IIC^{Man} homologues – 0.69 (dashed line in Fig. 3c). As noted above, the KLTEG sequence (alignment position 182–186 in Fig. 5a) that was proposed to contain an essential glutamate in IIC^{Man}, IIC^{Sor} and IIC^{Fru} (Lengeler *et al.*, 1994; Wehmeier *et al.*, 1995) is not

Fig. 4. Multiple alignment (a), mean hydropathy plot (b), and mean similarity plot (c) of the proteins comprising the IIC^{Man} family. Convention of presentation is as described in the legend to Fig. 3. Abbreviations and references to the published sequences are as described in Table 3. The mean hydropathy and similarity plots were determined for the four full-length proteins of this family, i.e. IIC^{Man}, IIC^{Fru}, IIC^{Sor} and IIC^{Aga}, but not for IIC^{Aga}. The proposed signature sequence of this IIC protein family is shown above the mean similarity plot.

conserved in the homologous IID^{Aga} which instead contains the MIARS sequence at the same alignment position. Also noteworthy is the cysteinyl residue at alignment position 37 (Fig. 5a) which is conserved in all three characterized IID proteins but not in IID^{Aga}. A recent report has shown that this cysteinyl is not essential for PTS activity since the C43E mutant of IID^{Man} retains

25 % of the PEP-dependent PTS activity as compared to the wild-type protein (Rhiel *et al.*, 1994).

The mean hydropathy plot, shown in Fig. 5(b), reveals that the N-termini of these proteins are fairly hydrophilic, whereas the central and C-terminal portions are strikingly hydrophobic. Using the Kyte & Doolittle (1982) method

Fig. 5. Multiple alignment (a), mean hydropathy plot (b), and mean similarity plot (c) of the proteins comprising the IID^{Man} family. Convention of presentation is as described in the legend to Fig. 3. The signature sequence of this IID^{Man} protein family is provided above the mean similarity plot. Abbreviations and references to the published sequences are as follows: IID^{Aga}(Eco), IID^{Aga} protein of *E. coli* (GenBank accession no. U18997); IID^{Fru}(Bsu), fructose-specific IID protein of *B. subtilis* (Martin-Verstraete *et al.*, 1990); IID^{Man}(Eco), mannose-specific IID protein of *E. coli* (previously named II-M^{Man}, Erni *et al.*, 1987); IID^{Sor}(Kpn), sorbose-specific IID protein of *K. pneumoniae* (Wehmeier & Lengeler, 1994; Wehmeier *et al.*, 1995).

and the TopPred II program (Claros & von Heijne, 1994), four to six helical segments were predicted for all of these proteins, and the corresponding N-terminal sequences were predicted to be localized to the cytoplasmic face of the membrane (data not shown). The signature sequence for this IID family is presented above the mean similarity plot shown in Fig. 5(c).

AgaX (ORF104)

The *aga* operon does not encode a IIA^{Man} homologue. However, screening the current databases (GenBank version 89.0; Swiss-Prot version 31.0) revealed an ORF (GenBank accession no. D26562; Fujita *et al.*, 1994) in the 2.4–4.1 min region of the *E. coli* chromosome which is

homologous to IIA^{Man}. Comparison scores of this ORF with IIA^{Man}, IIA^{Fru} and IIA^{Sor} were 10–13 SD (23–35% identity), thereby establishing that these four IIA proteins are homologous. IIA^{Aga} is 104 residues long whereas IIA^{Man} (136 residues), IIA^{Sor} (135 residues) and IIA^{Fru} (146 residues), are all substantially longer. The IIA^{Aga} protein proved to lack the N-terminal 40 residue segment which includes the reported active site histidyl phosphorylation site (His₁₀) in IIA^{Man} (Erni *et al.*, 1989). Although residue 11 in IIA^{Aga} is a histidyl residue, it is not surrounded by a region exhibiting significant sequence similarity to its protein homologues. It therefore seems unlikely that this truncated protein exhibits phosphoryl-transfer activity unless a sequencing error is responsible for the observed truncation.

Fig. 6. Evolutionary trees of the proteins or protein domains comprising the families of (a) IIA^{Man}, (b) IIB^{Man}, (c) IIC^{Man} and (d) IID^{Man}. The relative evolutionary distances (in arbitrary units) are shown adjacent to the branches. Construction of the trees was as described by Reizer & Reizer (1994) using the progressive alignment method of Feng & Doolittle (1990). Abbreviations used and references to the published sequences are described in the text and in the legends to Tables 2 and 3.

Phylogenetic relationships of Aga PTS proteins and their homologues

Fig. 6 presents the phylogenetic trees for the IIA, IIB, IIC and IID protein constituents of the Enzyme II^{Man}, Enzyme II^{Sor}, Enzyme II^{Fru}, and Enzyme II^{Aga} complexes. In all four trees the Aga proteins are more distant from their homologues than the latter proteins are from each other. When truncated proteins are present (IIA^{Aga} and IIC^{Aga}) these partial sequences exhibit long branch-lengths as expected. Only in the case of the IIB family

were the twin IIB^{Aga} proteins of comparable distance from their homologues. The similarities of the branching orders of the four trees, but the strikingly different relative branch lengths, are particularly worthy of note.

AgaA (ORF o163)

The *agaA* gene (see Fig. 1) encodes a protein of 163 amino acids that is homologous to two proteins in the current databases. One is the well-characterized *nagA* gene

Table 4. Binary comparisons of protein members of the glucosamine-6-phosphate-deaminase isomerase family

Values in parentheses below the designations of the proteins refer to the number of residues in that protein. The FASTA program using the dipeptide identities mode (ktup = 2) (Pearson & Lipman, 1988) was used to assess similarities of the indicated proteins. Values presented in the table which are not in brackets or parentheses represent percentage identities for segments having the numbers of compared residues indicated in parentheses. Comparison scores in standard deviations using the RDF2 program (Pearson & Lipman, 1988) and 150 shuffles are given in brackets below the values for percentage identity. The abbreviations used and the references to the published sequences are as follows: NagB, glucosamine-6-phosphate isomerase of *E. coli* [NagB(Eco), Rogers *et al.*, 1988]; AgaI, putative galactosamine-6-phosphate isomerase of *E. coli* [AgaI(Eco), GenBank accession no. U18997]; Nag1, glucosamine-6-phosphate isomerase of *Cand. albicans* [Nag1(Cal), Natarajan & Datta, 1993]; ORF f213, hypothetical 23.3 kDa protein encoded within the *tnaB-bglB* (81.5–84.5 min) region of the *E. coli* chromosome [ORF f213(Eco), Burland *et al.*, 1993]; ORF289, a putative protein of humans [ORF289(Hsa), N. Nomura and co-workers, GenBank accession no. D31766]; ORF126, a putative protein of *Myc. capricolum* [ORF126(Mca), Bork *et al.*, 1995; GenBank accession no. Z33055].

	NagB(Eco) (266)	Nag1(Cal) (248)	ORF f213(Eco) (213)	ORF289(Hsa) (289)	ORF126(Mca) (126; fragment)
AgaI(Eco) (251)	28 (201) [26]	25 (218) [24]	26 (203) [16]	27 (206) [30]	24 (108) [11]
NagB(Eco) (266)		47 (252) [76]	27 (212) [18]	60 (258) [101]	34 (126) [24]
Nag1(Cal) (248)			26 (220) [15]	47 (252) [62]	33 (126) [17]
ORF f213(Eco) (213)				28 (207) [17]	24 (90) [4]
ORF289(Hsa) (289)					35 (126) [23]

product of *E. coli*, the *N*-acetylglucosamine deacetylase (Plumbridge, 1989). The other is a functionally uncharacterized ORF from *Caenorhabditis elegans* (Wilson *et al.*, 1994). The *E. coli* enzyme is 382 residues long while the *Caen. elegans* protein is 418 residues long. The much shorter AgaA protein shows sequence similarity only with the C-terminal regions of its two homologues (data not shown). AgaA nevertheless exhibits 32–33 % identity in 111–114 overlapping residues (comparison scores of 16 sd) with its two homologues. Because the *agaA* gene is flanked in its upstream region by *agaW* with only 19 bp separating these two genes, the remainder of *agaA* cannot be present in this location. It is not clear whether *agaA* encodes a functional deacetylase (see Discussion).

AgaI (ORF o251)

Table 4 presents the binary comparison scores for AgaI with its homologues. AgaI is homologous to functionally well-characterized glucosamine-6-phosphate-deaminating isomerases from *E. coli* (NagB) and the yeast *Candida albicans* (Nag1), as well as to functionally uncharacterized *Mycoplasma capricolum*, *E. coli* and human ORFs. AgaI exhibits greatest similarity to the *E. coli* NagB protein (glucosamine-6-phosphate deaminase) and the human ORF289 (26–30 sd; see Table 4).

Recent chemical modification and site-directed mutagenesis studies performed with the purified glucosamine-

6-phosphate deaminase (NagB) of *E. coli* have shown that (a) two cysteinyl residues, Cys118 and Cys239, form a pair of vicinal thiols required for maximal catalytic activity and that (b) Cys239 plays a role in the allosteric activation of the enzyme by *N*-acetylglucosamine 6-phosphate, as well as in Zn²⁺ binding (Altamirano *et al.*, 1992). Multiple alignment of the proteins comprising the glucosamine-6-phosphate deaminase family showed that Cys118 and Cys231 of NagB are the only two fully conserved cysteinyls in AgaI, the human ORF289, and the yeast Nag1 proteins (Fig. 7a). These observations support the suggestion that AgaI shares similar mechanisms of catalysis and allosteric regulation with glucosamine-6-phosphate deaminases of *E. coli*, *Cand. albicans*, and humans. The mean similarity plot and phylogenetic tree for this family of proteins are shown in Fig. 7(b) and (c), respectively. The proposed signature sequence of this protein family is shown above the mean similarity plot (Fig. 7b).

AgaZ (ORF o426)

AgaZ (see Fig. 1) proved to be homologous to the protein encoded by *gatZ*, a functionally unidentified ORF (ORF378) in the galactitol (*gat*) operon (Nobelmann & Lengeler, 1995). These two proteins exhibit 53 % identity in 376 residues (comparison score of 131 sd). Screening of the current GenBank database (version 89.0) using the

Fig 7. Multiple alignment (a), mean similarity plot (b) and phylogenetic tree (c) of the proteins that comprise the glucosamine-6-phosphate-deaminase family. The proposed signature sequence of this protein family is shown above the mean similarity plot. Conventions of presentation are as described in the legends to Figs 3 and 6. Abbreviations used are as described in the legend to Table 4.

BLASTN program (Altschul *et al.*, 1990) revealed that a stretch of 78 nucleotides in the 3'-region of *agaZ* is nearly identical (96% identity) to a cDNA region immediately upstream of the homeobox-like-encoding sequence (*shoX*) of the marine sponge *Geodia cydonium* (Kruse *et al.*, 1994; Gamulin *et al.*, 1994). We expect that this cDNA sequence is artefactual. Although AgaZ and GatZ do not exhibit significant sequence similarity with other ORFs in the database, we suggest that they catalyse phosphorylation of tagatose 6-phosphate (see Discussion).

AgaY (ORF o286)

AgaY is homologous to proteins that comprise the metal-dependent aldolase (MDA) family, also referred to as the class II aldolases (Morse & Horecker, 1968; Marsh & Lebherz, 1992). As shown in Table 5, this family includes

four characterized yeast (*Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*) and bacterial (*E. coli* and *Corynebacterium glutamicum*) fructose-1,6-bisphosphate (FBP) aldolases which give high binary comparison scores with each other (60–164 sd) but exhibit much lower comparison scores (4–13 sd) with other members of the family. The other five protein members of the MDA family include two proteins of *Rhodobacter sphaeroides*, one of *B. subtilis* and the *E. coli* *gatY* and *agaY* gene products. The comparison scores obtained among this subset of aldolases (12–116 sd) are lower than those obtained among the four characterized FBP aldolases (see Table 5).

AgaY is most similar to GatY of *E. coli* (68 sd) and to the putative FBP aldolase of *B. subtilis* (63 sd). The comparison scores of AgaY with remaining members of the family are much lower (5–20 sd). The two proteins of *R. sphaeroides* are strikingly similar to each other (116 sd), but

Table 5. Binary comparisons of protein members of the MDA family

Values in parentheses below the designations of the proteins refer to the number of residues in that protein. The FASTA program using the dipeptide identities mode (ktup = 2) (Pearson & Lipman, 1988) was used to assess similarities of the indicated proteins. Values presented in the table which are not in brackets or parentheses represent percentage identities for segments having the number of compared residues indicated in parentheses. Comparison scores in standard deviations using the RDF2 program (Pearson & Lipman, 1988) and 150 shuffles are given in brackets below the values for percentage identity. The abbreviations used and the references to the published sequences are as follows: FBA (Sce), – (Spo), – (Eco), – (Cgl) and – (Bsu), fructose-1, 6-bisphosphate aldolases (FBAs) of *S. cerevisiae* (Schwelberger *et al.*, 1989; Rasmussen, 1994), *Schiz. pombe* (Mutoh & Hayashi, 1994), *E. coli* (Alefounder *et al.*, 1989; Alefounder & Perham, 1989), *Coryne. glutamicum* (von der Osten *et al.*, 1989) and *B. subtilis* (putative) (Trach *et al.*, 1988; Mitchell *et al.*, 1992), respectively; GatY (Eco) and AgaY (Eco), putative tagatose-1,6-bisphosphate aldolases that are encoded in the *gat* (Nobelmann & Lengeler, 1995) and *aga* operons of *E. coli* (GenBank accession no. U18997), respectively. The two putative sedoheptulose-1,7-bisphosphate aldolases (SBA) of *R. sphaeroides* [SBAI (Rsp), previously named CfxA, Gibson *et al.*, 1991; and SBAIL (Rsp), previously named CfxB, Chen *et al.*, 1991]. Thick lines segregate binary comparisons of (1) the characterized FBP aldolases with themselves (top left), (2) the putative tagatose-1,6-bisphosphate aldolases of *E. coli* and the *B. subtilis* aldolase (bottom right), and (3) the putative sedoheptulose-1,7-bisphosphate aldolases of *R. sphaeroides* (single central box).

	FBA (Spo) (358)	FBA (Eco) (359)	FBA (Cgl) (344)	SBAIL (Rsp) (354)	SBAI (Rsp) (359)	FBA (Bsu) (285)	AgaY (Eco) (286)	GatY (Eco) (286)
FBA (Sce) (359)	67 (360) [164]	49 (352) [127]	37 (347) [60]	33 (51) [11]	40 (43) [6]	33 (54) [7]	34 (44) [5]	29 (133) [11]
FBA (Spo) (358)		50 (357) [122]	37 (339) [77]	35 (49) [10]	36 (33) [5]	31 (55) [7]	41 (44) [7]	30 (111) [10]
FBA (Eco) (359)			40 (351) [63]	19 (243) [6]	15 (138) [6]	28 (193) [11]	34 (50) [6]	28 (165) [13]
FBA (Cgl) (344)				17 (321) [7]	16 (339) [4]	28 (167) [11]	25 (154) [13]	28 (47) [4]
SBAIL (Rsp) (354)					78 (356) [116]	25 (123) [14]	21 (135) [12]	34 (152) [17]
SBAI (Rsp) (359)						40 (156) [18]	31 (247) [20]	35 (203) [18]
FBA (Bsu) (285)							41 (286) [63]	42 (287) [64]
AgaY (Eco) (286)								54 (283) [68]

significantly less so to the other proteins in this family (4–20 SD; see Table 5). These intrafamily similarity scores clearly indicate that the nine proteins that comprise the metal-dependent aldolase family fall into three distinct clusters (see below).

Fig. 8 presents the multiple alignment of the MDA family revealing seven short regions of high sequence identity (indicated by roman numerals I–VII). All of these regions contain strongly polar residues that may be important for catalysis. Indeed, a recent study has identified the two conserved histidyl residues in region I, His-108 and His-111 of the *E. coli* FBP aldolase, as two of the zinc-binding ligands of this protein (Berry & Marshall, 1993). The concomitant loss of FBP aldolase activity and zinc content in the mutant H108A and H111A proteins confirms the essential role of zinc in catalysis by the *E. coli* FBP aldolase. Replacement of the partially conserved residues Cys-112 and His-142 with alanine resulted in partial loss of zinc binding and enzyme activity (Berry & Marshall, 1993). The lower zinc content of these mutant proteins may be due to the close proximity of Cys-112 and His-142

to the residues that are directly involved in zinc co-ordination. Based on these observations and the multiple alignment shown in Fig. 8, we propose that the remaining histidyl, glutamyl and/or aspartyl residue(s) which play(s) a role in zinc chelation is(are) common to all nine sequenced class II aldolases (see Fig. 8 regions II, III, IV, V and VII). Also noteworthy are the mutations of the putative FBP aldolase of *B. subtilis*, P205L (two residues preceding region V; see Fig. 8) and V232E (region VI), as well as the FBP aldolase mutant of *E. coli*, V300G (10 residues downstream of region VI) which result in thermolability of the enzyme (Mitchell *et al.*, 1992; Singer *et al.*, 1991a, b). Interestingly, all of these mutants are conditionally lethal as the mutations preferentially inhibit stable RNA (but not mRNA) synthesis upon shift to the non-permissive temperature (see Discussion).

The conserved regions shown in the multiple alignment of the MDA proteins were used to construct two signature sequences that proved to be specific to currently recognized members of this family. The first signature sequence, [LIVM]EXE[LIVM]GX₂G[GSTA]XE, corresponds to

<

Fig 8. Multiple alignment of the proteins comprising the MDA family. Residues conserved in all nine sequenced proteins are boxed. Residues common in at least five of the sequenced proteins are presented in the consensus sequence (consensus). Numbers to the left of the sequences shown correspond to the residue numbers in the individual proteins. Numbers above the aligned sequences refer to the alignment position and not to the residue number in any one of the aligned proteins. Abbreviations and references to published sequences are as described in the legend to Table 5. The seven regions of strong conservation in all members of this protein family are indicated by roman numerals I–VII.

Fig. 9. Phylogenetic tree of proteins of the MDA family. Convention of presentation is as described in Fig. 6. Abbreviations and references to the published sequences are as described in the legend to Table 5.

region III in the multiple alignment and contains three conserved acidic residues. The second signature sequence, [LIVM]AX[STAG]XGX₂HGXY, corresponds to conserved region IV of these proteins (see Fig. 8).

Fig. 9 presents the phylogenetic tree for the MDA family. As suggested from considerations of the binary comparison scores reported in Table 5, three clusters are discernible. The first loose cluster includes the four characterized FBP aldolases of *E. coli*, yeast, and *Coryne. glutamicum* (right-hand side in Fig. 9). Surprisingly, the *E. coli* protein is closer to the two yeast aldolases than to the *Coryne. glutamicum* enzyme. This apparent incongruency between the 'protein tree' and the expected 'species tree' was suggested to be due to horizontal transfer of a class II aldolase gene from some eubacterium to yeast (Smith *et al.*, 1992). The second cluster includes the two enzymes from *R. sphaeroides* which cluster tightly together (lower left-hand side in Fig. 9). These two proteins are respectively encoded within the form I and form II Calvin cycle (CO₂ fixation) operons which are found on two distinct loci of the chromosome of this organism. Although these enzymes have been proposed to be FBP aldolases (Gibson *et al.*, 1991; Chen *et al.*, 1991), we suggest, on the basis of their clustering pattern in Fig. 9 and their occurrence in CO₂-fixation operons, that they exhibit specificity for sedoheptulose 1,7-bisphosphate. Finally, the three proteins, AgaY(Eco), GatY(Eco) and FBA(Bsu), comprise the third, somewhat loose cluster. Because of the occurrence of GatY in the galactitol operon which encodes ketose bisphosphate aldolase (Nobelmann & Lengeler, 1995), we suggest that GatY, as well as AgaY are tagatose-1,6-bisphosphate aldolases (see Discussion).

Neither AgaY nor GatY exhibits significant similarity with the sequenced tagatose-1,6-bisphosphate aldolases which are encoded within the *lac* operons of *Streptococcus*

mutans (Rosey & Stewart, 1992), *L. lactis* (de Vos *et al.*, 1990) and *Staphylococcus aureus* (Rosey *et al.*, 1991). The tagatose-1,6-bisphosphate aldolases of these Gram-positive bacteria comprise a distinct family of proteins lacking apparent sequence similarity with hitherto characterized ketose bisphosphate aldolases (Rosey & Stewart, 1992; unpublished data).

AgaS (ORF o384)

When the AgaS sequence was screened against the current databanks with the BLAST program (employing the Blosum62 matrix), the highest similarity, albeit of doubtful significance, was detected with a putative glutamine:fructose-6-phosphate amidotransferase (trivial name glucosamine-6-phosphate synthase) of *Mycobacterium leprae* (probability of matching by chance $P = 0.28$). Using the SSEARCH and RSS programs, which are included in the FASTA program package (Pearson, 1994), subsequent sequence comparisons between AgaS and glucosamine-6-phosphate synthases revealed that AgaS is homologous to the C-terminal region of characterized and putative glucosamine-6-phosphate synthases of *E. coli* (the *glmS* gene product; Swiss-Prot accession no. P17169; 19% identity in 336 amino acids overlap; 9 sd), of *Myc. leprae* (Swiss-Prot accession no. P40831; 19% in 302 residues overlap; 10 sd), and of *Rhizobium mililoti* and *Rhizobium leguminosarum* (the *nodM* gene protein products; Swiss-Prot accession nos P25195 and P08633, respectively; 21% identity in 140–143 amino acid overlap; 8–9 sd). Lower similarity scores (1–5 sd) were obtained when AgaS was compared to glucosamine-6-phosphate synthases of eukaryotic organisms, i.e. *S. cerevisiae* and humans. We note that the *E. coli* GlmS protein was shown to consist of two independently folding structural domains that catalyse two coupled enzymic reactions: N-terminal domain (240 residues) which binds glutamine, and the C-terminal domain (368 residues) which binds fructose 6-phosphate and bear the ketose–aldose isomerase activity (Denisot *et al.*, 1991). Interestingly, the region of similarity between AgaS and GlmS starts at the hinge region that connects the two domains (Denisot *et al.*, 1991), and it encompasses the entire ketose–aldose-isomerase-bearing domain (data not shown).

DISCUSSION

E. coli is capable of *N*-acetylgalactosamine utilization (unpublished observations), but the genes and enzymes responsible for its metabolism have never been characterized. The most probable pathway for the utilization of *N*-acetylgalactosamine (AGA) in *E. coli* is shown in Fig. 10. According to this scheme, exogenous AGA is transported and phosphorylated by the PTS and then metabolized in a pathway that parallels that for *N*-acetylglucosamine metabolism (see Fig. 10). The intracellular phosphorylated acetylated sugar is first deacetylated and then deaminated, resulting in tagatose 6-phosphate as expected for a sugar of the galacto configuration. This sugar is then phosphorylated by an ATP-dependent kinase to give tagatose 1,6-bisphosphate which

Fig. 10. Schematic presentation of the proposed metabolic pathway for *N*-acetylgalactosamine and the characterized parallel pathway for *N*-acetylglucosamine. The gene designations corresponding to the enzymes catalysing or postulated to catalyse each step are indicated below the arrows.

is cleaved by an aldolase to two triose phosphates. We postulate that the PTS-catalysed step 1 involves AgaX or ManX, AgaB, AgaB', AgaC and AgaD (all homologues of Enzyme II^{Man} constituents), step 2 is catalysed by AgaA (a truncated version of NagA), step 3 is catalysed by AgaI (homologous to NagB), step 4 may be catalysed by AgaZ (homologous to GatZ) and step 5 may be catalysed by AgaY (homologous to GatY). Thus the *aga* genes analysed here potentially provide the complete pathway for the degradation of AGA to triose phosphates, the first common intermediates in the glycolytic pathway. The striking sequence similarities noted between AgaZ and GatZ as well as AgaY and GatY (53 and 54% identity, respectively) clearly suggest that these pairs of enzymes serve the same biochemical functions. This postulated pathway accounts for all but one of the genes present in the *aga* gene cluster. The *agaS* gene encodes a protein homologous to the C-terminal domain that bears the fructose-6-phosphate-binding site and the aldose-ketose isomerase function of glutamine:fructose-6-phosphate amidotransferase (glucosamine-6-phosphate synthase). By analogy, *agaS* may encode a tagatose-6-phosphate-binding protein that bears a ketose-aldose isomerase activity. We are not aware of a glutamine:tagatose-6-phosphate amidotransferase (galactosamine-6-phosphate synthase) activity in *E. coli*. Nevertheless, it is possible that a glutamine-binding domain, homologous to the N-terminal portion of GlnS, is encoded outside the *aga* gene cluster, and that these two proteins comprise the postulated synthase. In this regard it is interesting to note that *glmS* and the *nag* genes of *E. coli* appear to constitute a single regulon (Plumbridge *et al.*, 1993). Thus, the presence of *agaS* in the *aga* gene cluster, may represent a parallel situation.

In addition to *N*-acetylgalactosamine, *N*-acetylmannosamine can be utilized as a carbon source for growth of enteric bacteria (Gutnick *et al.*, 1969). The *aga* gene cluster can therefore be considered to be involved in *N*-acetylmannosamine metabolism rather than, or in addition to, *N*-acetylgalactosamine metabolism. We consider this possibility less likely because of the high degree of sequence similarity between AgaY and GatY (which are 54% identical) as well as between AgaZ and GatZ (which are 53% identical). These pairs of enzymes are proposed to be the aldolase and kinase, respectively,

which may act on sugars of the same (i.e. galacto) configuration.

The PTS proteins of the Enzyme II^{Aga} complex (IIA, IIB, IIB', IIC, IIC' and IID) proved to be homologous to the previously established 'splinter' group of Enzymes II of the PTS (Reizer *et al.*, 1991b), which includes the mannose Enzyme II of *E. coli*, the fructose Enzyme II of *B. subtilis* and the recently sequenced sorbose Enzyme II of *K. pneumoniae* (Nobelmann & Lengeler, 1995; Wehmeier *et al.*, 1995). These enzyme complexes exhibit little sequence similarity with other sequenced Enzyme II proteins of the PTS which comprise the major class of the PTS permeases. We believe that the proteins which comprise the 'splinter' group of PTS Enzyme II complexes evolved independently of the others.

The Aga proteins encoded within the *aga* gene cluster include two complete IIB proteins, one complete and one truncated IIC protein, and a single IID protein. Although no IIA protein was found in the *aga* gene cluster, the *agaX* gene, found in the 2.4–4.1 min region of the *E. coli* chromosome proved to be homologous to IIA^{Man}. It should be noted, however, that AgaX appears to lack the N-terminal region that includes the phosphorylation site in the other IIA proteins of this family. As a result, it cannot be concluded that AgaX exhibits catalytic function. In this regard, it is interesting to note recent data suggesting that the phosphoryltransfer pathway from HPr to IIB^{Man} can bypass (albeit at a slow rate) the IIA^{Man} protein domain as an obligatory phosphoryltransfer intermediate (Stolz *et al.*, 1993). It is possible either that IIA^{Man} phosphorylates IIB^{Aga} and/or IIB^{Aga'} or that the IIA function is not required. The truncated IIA^{Aga} identified here may be a non-functional remnant of a functional IIA^{Aga}. In this regard, it is relevant to note that IIA^{Glc} of the major PTS permease cluster phosphorylates IIB^{Glc}, IIB^{Ser}, IIB^{Tre} and IIB^{Mal} (Postma *et al.*, 1993).

Two Aga PTS proteins (IIC and IID) are integral constituents of the membrane. Topological analyses indicate that these proteins probably span the membrane 7 and 4–6 times, respectively. These predictions, based on hydropathy analyses of the multiply aligned and individual members of the IIC and IID families, are at odds with previous predictions (Lengeler *et al.*, 1994).

The putative deacetylase (AgaA) is an N-terminally truncated homologue of the *E. coli* N-acetylglucosamine-6-phosphate deacetylase (NagA). Because the N-terminal truncation eliminates much of the region that is conserved in the NagA protein of *E. coli* and the putative NagA protein of *Caen. elegans*, it cannot be concluded that AgaA is functional for catalysis. However, AgaI (a putative galactosamine-6-phosphate-isomerizing deaminase) proved to be homologous throughout its length with several other deaminases including the well-characterized *E. coli* glucosamine-6-phosphate deaminase (NagB), and therefore is probably catalytically active.

We have suggested that AgaZ and the homologous *E. coli* GatZ (of unknown function) are both ATP:tagatose-6-phosphate kinases. They do not, however, exhibit significant sequence similarity to other sugar kinases in the current databases, including the sequenced tagatose-6-phosphate kinases of *L. lactis* (de Vos *et al.*, 1990), *Strep. mutans* (Rosey & Stewart, 1992) and *Staph. aureus* (Rosey *et al.*, 1991). In this regard it is important to note that currently sequenced sugar kinases comprise several distinct families rather than a coherent family (Reizer *et al.*, 1991a; Bork *et al.*, 1993). It may also be relevant that the characterized tagatose-1,6-bisphosphate aldolases sequenced from these Gram-positive bacteria are not homologous to sequenced ketose bisphosphate aldolases of the MDA family (see Results).

The data presented here and in a recent study (Nobelman & Lengeler, 1995) suggest that two distinct but otherwise homologous tagatose-1,6-bisphosphate aldolases, GatY and AgaY, are encoded within the *E. coli* *gat* and *aga* operons which map on the chromosome at 46.7 and 70.5 min, respectively. An early study showed that a temperature-sensitive ketose bisphosphate aldolase, encoded by the *kba* gene within the 71 min region, rather than GatY, assumes the role of tagatose 1,6-bisphosphate cleavage in galactitol metabolism (Lengeler, 1977). Because *agaY* is the only recognizable aldolase-encoding gene within the 71 min region of the *E. coli* chromosome (unpublished observations) we conclude that the previously designated *kba* gene is in fact *agaY*. The preferential use of AgaY rather than GatY during galactitol metabolism is conceivably due to an as-yet-unknown mutation in the gene encoding GatY. Such 'crosstalk' between operons may allow completion of the N-acetylglactosamine catabolic pathway using enzymes encoded outside of the *aga* gene cluster (i.e. IIA of the PTS and the deacetylase).

AgaY is homologous to MDAs which, according to the phylogenetic tree portrayed in Fig. 9, fall into three principal clusters. These we have suggested correlate with their sugar specificity, being specific for fructose 1,6-bisphosphate, tagatose 1,6-bisphosphate, and sedoheptulose 1,7-bisphosphate. In all cases, these aldolases recognize triose phosphates (or erythrose 4-phosphate in the case of the sedoheptulose-1,7-bisphosphate aldolase), all of which can be metabolized via known pathways. Since these metabolites can be acted upon by ubiquitous enzymes, AgaY represents the last substrate-specific

enzyme required for the degradation of N-acetylglactosamine.

Recent studies have demonstrated that stable RNA synthesis by the *rrnB* operon is strongly inhibited upon shift to the non-permissive temperature (42 °C) in *E. coli* mutants bearing a thermolabile FBP aldolase (*ts8* and *b8*) (Singer *et al.*, 1991a, b). Inhibition was shown to occur at the level of transcriptional initiation and appeared to be due to elevated levels of FBP derived from glycolytic carbohydrates. Thus, transcription from the *rrnBP1* and *rrnBP2* promoters was refractory to thermal inactivation of FBP aldolase when these mutants were grown at the non-permissive temperature on carbon sources which enter the central carbon-metabolic pathways past FBP (i.e. glycerol, pyruvate or succinate) or when exposed to the non-metabolizable glucose analogue 2-deoxyglucose, following growth at 42 °C in glycerol-containing minimal medium (Singer *et al.*, 1991b). These observations lead us to propose, as a guide for future studies, that a metabolic mechanism involving the *trans*-acting global transcriptional regulatory protein FruR, and its effector molecule(s) fructose 1-phosphate and/or FBP, play a role in regulating expression of stable RNA-encoding genes. In support of this proposal, we have identified a potential FruR binding site (5'-TACCGTTTCACT-3') upstream of the *E. coli* *rrnB* operon. Since rRNA synthesis in mutants of *B. subtilis* bearing a thermosensitive FBP aldolase is also temperature-sensitive (L. Brown, unpublished data cited in Trach *et al.*, 1988, and in Mitchell *et al.*, 1992), we further propose that transcriptional regulation of stable RNA-encoding genes in *B. subtilis* is mediated by a FruR analogue in a manner similar to that proposed for *E. coli*.

In summary, we have provided compelling circumstantial evidence for the proposal that the *aga* gene cluster functions in the catabolism of N-acetylglactosamine. The pathway is probably not, however, encoded entirely by genes within the *aga* gene cluster. AgaX or the IIA domain of ManX (Erni *et al.*, 1987), may provide the IIA^{Aga} function. Similarly, the deacetylase encoded within the *aga* gene cluster is truncated, and may be non functional. If so, the deacetylase function may be provided by a deacetylase encoded outside of the *aga* gene cluster.

The Aga PTS is unusual in three respects: (1) no IIA protein is encoded within the *aga* gene cluster; (2) there are two IIB proteins, both of which may be functional for phosphoryl transfer; (3) there are two IIC proteins, one of which is full length, the other of which is truncated. The mannitol-, glucose- and fructose-specific Enzyme II complexes have been shown to form oligomers, most likely dimers (Leonard & Saier, 1983; Saier & Leonard, 1983; Erni, 1986; Jacobson, 1992; Boer *et al.*, 1994; Charbit *et al.*, 1996). The same may be true of the *E. coli* mannose PTS permease (Erni *et al.*, 1987; Saier & Reizer, 1992; Stolz *et al.*, 1993). The IIB^{Aga} and IIB^{Aga'} proteins may together form a heterooligomeric complex, and the same may be true of the IIC^{Aga} and IIC^{Aga'} proteins, even though the latter protein is truncated. It should be noted in this regard that artificial truncation of the fructose permease of *Xanthomonas campestris* does not prevent its

association with the wild-type protein (de Crécy-Lagard *et al.*, 1991). Further studies will be required to establish the novel aspects of structure and function in the putative N-acetylgalactosamine Enzyme II complex of the PTS.

ACKNOWLEDGEMENTS

This work was supported by US Public Health Service grants 5R01AI21702 and 2R01AI14176 from the National Institute of Allergy and Infectious Diseases.

NOTE ADDED IN PROOF

Recent sequence analyses revealed that the AgaX protein (accession no. P36881) has a putative N-terminal extension of 42 residues (K. Rudd, personal communication). The extended protein (146 residues; 16540 Da) is similar in length to IIA^{Man}, IIA^{Sor} and IIA^{Fru} (135–146 residues) and it contains a histidyl residue (position 9) which aligns with the active site phosphorylatable histidyl residues of the homologous IIA proteins. It therefore seems likely that AgaX exhibits phosphoryl transfer activity.

REFERENCES

- Alefunder, P. R. & Perham, R. N. (1989). Identification, molecular cloning and sequence analysis of a gene cluster encoding the class II fructose 1,6-bisphosphate aldolase, 3-phosphoglycerate kinase and a putative second glyceraldehyde 3-phosphate dehydrogenase of *Escherichia coli*. *Mol Microbiol* **3**, 723–732.
- Alefunder, P. R., Baldwin, D. A., Perham, R. N. & Short, N. J. (1989). Cloning, sequence analysis and over-expression of the gene for the class II fructose 1,6-bisphosphate aldolase of *Escherichia coli*. *Biochem J* **257**, 529–534.
- Altamirano, M. M., Plumbridge, J. A. & Calcagno, M. L. (1992). Identification of two cysteine residues forming a pair of vicinal thiols in glucosamine-6-phosphate deaminase from *Escherichia coli* and a study of their functional role by site-directed mutagenesis. *Biochemistry* **31**, 1153–1158.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol* **215**, 403–410.
- von Bodman, B. S., Hayman, G. T. & Farrand, S. K. (1992). Opine catabolism and conjugal transfer of the nopaline Ti plasmid pTiC58 are coordinately regulated by a single repressor. *Proc Natl Acad Sci USA* **89**, 643–647.
- Berry, A. & Marshall, K. E. (1993). Identification of zinc-binding ligands in the class II fructose-1,6-bisphosphate aldolase of *Escherichia coli*. *FEBS Lett* **318**, 11–16.
- Boer, H., ten Hoeve-Duurkens, R. H., Schuurman-Wolters, G. K., Dijkstra, A. & Robillard, G. T. (1994). Expression, purification, and kinetic characterization of the mannitol transport domain of the phosphoenolpyruvate-dependent mannitol phosphotransferase system of *Escherichia coli*. *J Biol Chem* **269**, 17863–17871.
- 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.
- Bork, P., Ouzounis, C., Casari, G., Schneider, R., Sander, C., Dolan, M., Gilbert, W. & Gillevet, P. M. (1995). Exploring the *Mycoplasma capricolum* genome: a minimal cell reveals its physiology. *Mol Microbiol* **16**, 955–967.
- Burland, V. D., Plunkett, G., III, Daniels, D. L. & Blattner, F. R. (1993). DNA sequence and analysis of 136 kilobases of the *Escherichia coli* genome: organizational symmetry around the origin of replication. *Genomics* **16**, 551–561.
- Burland, V., Plunkett, G., III, Sofia, H. J., Daniels, D. L. & Blattner, F. R. (1995). Analysis of the *Escherichia coli* genome VI: DNA sequence of the region from 92.8 through 100 minutes. *Nucleic Acids Res* **23**, 2105–2119.
- Claros, M. G. & von Heijne, G. (1994). TopPred II – an improved software for membrane protein structure predictions. *Comp Appl Biosci* **10**, 685–686.
- Charbit, A., Reizer, J. & Saier, M. H., Jr (1996). Function of the duplicated IIB domain and oligomeric structure of the fructose permease of *E. coli*. *J Biol Chem* (in press).
- Chen, J. H., Gibson, J. L., McCue, L. A. & Tabita, F. R. (1991). Identification, expression, and deduced primary structure of transketolase and other enzymes encoded within the form II CO₂ fixation operon of *Rhodobacter sphaeroides*. *J Biol Chem* **266**, 20447–20452.
- Choi, Y. L., Kawase, S., Nishida, T., Sakai, H., Komano, T., Kawamukai, M., Utsumi, R., Kohara, Y. & Akiyama, K. (1988). Nucleotide sequence of the *glpR* gene encoding the repressor for the glycerol-3-phosphate regulon of *Escherichia coli* K12. *Nucleic Acids Res* **16**, 7732.
- de Crécy-Lagard, V., Bouvet, O. M. M., Lejeune, P. & Danchin, A. (1991). Fructose catabolism in *Xanthomonas campestris* pv. *campestris*. Sequence of the PTS operon, characterization of the fructose-specific enzymes. *J Biol Chem* **266**, 18154–18161.
- Denisot, M.-A., Le Goffic, F. & Badet, B. (1991). Glucosamine-6-phosphate synthase from *Escherichia coli* yields two proteins upon limited proteolysis: identification of the glutamine amidohydrolase and 2R ketose/aldose isomerase-bearing domains based on their biochemical properties. *Arch Biochem Biophys* **288**, 225–230.
- Devereux, J., Haeberli, P. & Smithies, O. (1984). A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res* **12**, 387–395.
- Erni, B. (1986). Glucose-specific permease of the bacterial phosphotransferase system: phosphorylation and oligomeric structure of the glucose-specific II^{Glc}-III^{Glc} complex of *Salmonella typhimurium*. *Biochemistry* **25**, 312–318.
- Erni, B. (1992). Group translocation of glucose and other carbohydrates by the bacterial phosphotransferase system. *Int Rev Cytol* **137**, 127–148.
- Erni, B., Zanolari, B. & Kocher, H. P. (1987). The mannose permease of *Escherichia coli* consists of three different proteins. Amino acid sequence and function in sugar transport, sugar phosphorylation, and penetration of phage lambda DNA. *J Biol Chem* **262**, 5238–5247.
- Erni, B., Zanolari, B., Graff, P. & Kocher, H. P. (1989). Mannose permease of *Escherichia coli*. Domain structure and function of the phosphorylating subunit. *J Biol Chem* **264**, 18733–18741.
- Feng, D.-F. & Doolittle, R. F. (1990). Progressive alignment and phylogenetic tree construction of protein sequences. *Methods Enzymol* **183**, 375–387.
- Fujita, N., Mori, H., Yura, T. & Ishihama, A. (1994). Systematic sequencing of the *Escherichia coli* genome: analysis of the 2.4–4.1 min (110,917–193,643 bp) region. *Nucleic Acids Res* **22**, 1637–1639.
- Gamulin, V., Rinkevich, B., Schäcke, H., Kruse, M., Müller, I. M. & Müller, W. E. G. (1994). Cell adhesion receptors and nuclear receptors are highly conserved from the lowest metazoa (marine sponges) to vertebrates. *Biol Chem Hoppe-Seyler* **375**, 583–588.
- Gibson, J. L., Falcone, D. L. & Tabita, F. R. (1991). Nucleotide sequence, transcriptional analysis, and expression of genes encoded

- within the form I CO₂ fixation operon of *Rhodobacter spheroids*. *J Biol Chem* **266**, 14646–14653.
- Gutnick, D., Calvo, J. M., Klopotoski, T. & Ames, B. N. (1969). Compounds which serve as the sole source of carbon or nitrogen for *Salmonella typhimurium* LT-2. *J Bacteriol* **100**, 215–219.
- von Heijne, G. (1994). Membrane proteins – from sequence to structure. *Annu Rev Biophys Biomol Struct* **23**, 167–192.
- Jacobson, G. R. (1992). Interrelationship between protein phosphorylation and oligomerization in transport and chemotaxis via the *Escherichia coli* mannitol phosphotransferase system. *Res Microbiol* **143**, 113–116.
- Jacobson, G. R. & Saraceni-Richards, C. (1993). The *Escherichia coli* mannitol permease as a model for transport via the bacterial phosphotransferase system. *J Bioenerg Biomembr* **25**, 621–626.
- Komoda, Y., Enomoto, M. & Tominaga, A. (1991). Large inversion in *Escherichia coli* K-12 1485IN between inversely oriented IS3 elements near *lac* and *cdd*. *Genetics* **129**, 639–645.
- Kruse, M., Mikoc, A., Cetkovic, H., Gamulin, V., Rinkevich, B., Müller, I. M. & Müller, W. E. G. (1994). Cloning of the homeo-domain from the marine sponge *Geodia cydonium*. *Mech Ageing Dev* **77**, 43–54.
- Kyte, J. & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J Mol Biol* **157**, 105–132.
- Lengeler, J. (1977). Analysis of mutations affecting the dissimilation of galactitol (dulcitol) in *Escherichia coli* K12. *Mol & Gen Genet* **152**, 83–91.
- Lengeler, J. W., Jahreis, K. & Wehmeier, U. F. (1994). Enzymes II of the phosphoenolpyruvate-dependent phosphotransferase systems: their structure and function in carbohydrate transport. *Biochim Biophys Acta* **1188**, 1–28.
- Leonard, J. E. & Saier, M. H., Jr (1983). Mannitol-specific Enzyme II of the bacterial phosphotransferase system II. Reconstitution of vectorial transphosphorylation in phospholipid vesicles. *J Biol Chem* **258**, 10757–10760.
- Lu, Z. & Lin, E. C. (1989). The nucleotide sequence of *Escherichia coli* genes for L-fucose dissimilation. *Nucleic Acids Res* **17**, 4883–4884.
- Marsh, J. J. & Leberherz, H. G. (1992). Fructose-bisphosphate aldolases: an evolutionary history. *Trends Biochem Sci* **17**, 110–113.
- Martin-Verstraete, I., Debarbouille, M., Klier, A. & Rapoport, G. (1990). Levanase operon of *Bacillus subtilis* includes a fructose-specific phosphotransferase system regulating the expression of the operon. *J Mol Biol* **214**, 657–671.
- Meins, M., Jenö, P., Müller, D., Richter, W. J., Rosenbusch, J. P. & Erni, B. (1993). Cysteine phosphorylation of the glucose transporter of *Escherichia coli*. *J Biol Chem* **268**, 11604–11609.
- Mitchell, C., Morris, P. W., Lum, L., Spiegelman, G. & Vary, J. C. (1992). The amino acid sequence of a *Bacillus subtilis* phosphoprotein that matches an *orfY*–*tsr* coding sequence. *Mol Microbiol* **6**, 1345–1349.
- Morse, D. E. & Horecker, B. L. (1968). The mechanism of action of aldolases. *Adv Enzymol Relat Areas Mol Biol* **31**, 125–181.
- Mutoh, N. & Hayashi, Y. (1994). Molecular cloning and nucleotide sequencing of *Schizosaccharomyces pombe* homologue of the class II fructose-1,6-bisphosphate aldolase gene. *Biochim Biophys Acta* **1183**, 550–552.
- Natarajan, K. & Datta, A. (1993). Molecular cloning and analysis of the NAG1 cDNA coding for glucosamine-6-phosphate deaminase from *Candida albicans*. *J Biol Chem* **268**, 9206–9214.
- Nobelmann, B. & Lengeler, J. W. (1995). Sequence of the *gat* operon for galactitol utilization from wild-type strain EC3132 of *Escherichia coli*. *Biochim Biophys Acta* **1262**, 69–72.
- Oskouian, B. & Stewart, G. C. (1990). Repression and catabolite repression of the lactose operon of *Staphylococcus aureus*. *J Bacteriol* **172**, 3804–3812.
- von der Osten, C. H., Barbas, C. F., III, Wong, C.-H. & Sinskey, A. J. (1989). Molecular cloning, nucleotide sequence and fine-structural analysis of the *Corynebacterium glutamicum* *fda* gene: structural comparison of *C. glutamicum* fructose-1,6-bisphosphate aldolase to class I and class II aldolases. *Mol Microbiol* **3**, 1625–1637.
- Pas, H. H., Meyer, G. H., Kruizinga, W. H., Tamminga, K. S., van Weeghel, R. P. & Robillard, G. T. (1991). Phospho-NMR demonstration of phosphocysteine as a catalytic intermediate on the *Escherichia coli* phosphotransferase system EII^{MAN}. *J Biol Chem* **266**, 6690–6692.
- Pearson, W. R. (1994). Using the FASTA program to search protein and DNA sequence databases. In *Methods in Molecular Biology: Computer Analysis of Sequence Data*, part 2, pp. 365–389. Edited by A. M. Griffin & H. G. Griffin. Totowa, NJ: Humana Press.
- Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* **85**, 2444–2448.
- Plumbridge, J. A. (1989). Sequence of the *nagBACD* operon in *Escherichia coli* K12 and pattern of transcription within the *nag* regulon. *Mol Microbiol* **3**, 505–515.
- Plumbridge, J. A., Cochet, O., Souza, J. M., Altamirano, M. M., Calcagno, M. L. & Badet, B. (1993). Coordinated regulation of amino sugar-synthesizing and -degrading enzymes in *Escherichia coli* K-12. *J Bacteriol* **175**, 4951–4956.
- Plunkett, G., III, Burland, V., Daniels, D. L. & Blattner, F. R. (1993). Analysis of the *Escherichia coli* genome. III. DNA sequence of the region from 87.2 to 89.2 minutes. *Nucleic Acids Res* **21**, 3391–3398.
- Postma, P., Lengeler, J. & Jacobson, G. R. (1993). Phosphoenolpyruvate:carbohydrate phosphotransferase systems of bacteria. *Microbiol Rev* **57**, 543–594.
- Rasmussen, S. W. (1994). Sequence of a 28.6 kb region of yeast chromosome XI includes the FBA1 and TOA2 genes, an open reading frame (ORF) similar to a translationally controlled tumor protein, one ORF containing motifs also found in plant storage proteins and 13 ORFs with weak or no homology to known proteins. *Yeast* **10**, S63–68.
- Reizer, A. & Reizer, J. (1994). Progressive multiple alignment of protein sequences and the construction of phylogenetic trees. In *Methods in Molecular Biology: Computer Analysis of Sequence Data*, part 2, pp. 319–325. Edited by A. M. Griffin & H. G. Griffin. Totowa, NJ: Humana Press.
- Reizer, A., Deutscher, J., Saier, M. H., Jr & Reizer, J. (1991a). Analysis of the gluconate (*gnt*) operon of *Bacillus subtilis*. *Mol Microbiol* **5**, 1081–1089.
- Reizer, A., Pao, G. M. & Saier, M. H., Jr (1991b). Evolutionary relationships among the permease proteins of the bacterial phosphoenolpyruvate:sugar phosphotransferase system. Construction of phylogenetic trees and possible relatedness to proteins of eukaryotic mitochondria. *J Mol Evol* **33**, 179–193.
- Reizer, J., Michotey, V., Reizer, A. & Saier, M. H., Jr (1994a). Novel phosphotransferase system genes revealed by bacterial genome analysis: unique, putative fructose- and glucoside-specific systems. *Prot Sci* **3**, 3440–3450.
- Reizer, J., Reizer, A., Kornberg, H. L. & Saier, M. H., Jr (1994b). Sequence of the *fruB* gene of *Escherichia coli* encoding the diphosphoryl transfer protein (DTP) of the phosphoenolpyruvate:sugar phosphotransferase system. *FEMS Microbiol Lett* **118**, 159–162.
- Reizer, J., Reizer, A. & Saier, M. H., Jr (1995). Novel phosphot-

ransferase system genes revealed by bacterial genome analysis – a gene cluster encoding a unique Enzyme I and the proteins of a fructose-like permease system. *Microbiology* **141**, 961–971.

Rephaeli, A. W. & Saier, M. H., Jr (1980). Substrate specificity and kinetic characterization of sugar uptake and phosphorylation, catalyzed by the mannose enzyme II of the phosphotransferase system in *Salmonella typhimurium*. *J Biol Chem* **255**, 8585–8591.

Rhiel, E., Flükiger, K., Wehrli, C. & Erni, B. (1994). The mannose transporter of *Escherichia coli* K12, oligomeric structure, and function of two conserved cysteines. *Biol Chem Hoppe-Seyler* **375**, 551–559.

Rogers, M. J., Ohgi, T., Plumbridge, J. & Söll, D. (1988). Nucleotide sequences of the *Escherichia coli* *nagE* and *nagB* genes: the structural genes for the *N*-acetylglucosamine transport protein of the bacterial phosphoenolpyruvate:sugar phosphotransferase system and for glucosamine-6-phosphate deaminase. *Gene* **62**, 197–207.

van Rooijen, R. J., Dechering, K. J., Niek, C., Wilmink, J. & de Vos, W. M. (1993). Lysines 72, 80 and 213 and aspartic acid 210 of the *Lactococcus lactis* LacR repressor are involved in the response to the inducer tagatose-6-phosphate leading to induction of *lac* operon expression. *Prot Eng* **6**, 201–206.

Rosey, E. L. & Stewart, G. C. (1992). Nucleotide and deduced amino acid sequences of the *lacR*, *lacABCD*, and *lacFE* genes encoding the repressor, tagatose-6-phosphate gene cluster, and sugar-specific PTS components of the lactose operon of *Streptococcus mutans*. *J Bacteriol* **174**, 6159–6170.

Rosey, E. L., Oskouian, B. & Stewart, G. C. (1991). Lactose metabolism by *Staphylococcus aureus*: characterization of *lacABCD*, the structural genes of the tagatose-6-phosphate pathway. *J Bacteriol* **173**, 5992–5998.

Saier, M. H., Jr & Leonard, J. E. (1983). The mannitol enzyme II of the bacterial phosphotransferase system: a functionally chimaeric protein with receptor, transport, kinase and regulatory activities. In *Multifunctional Proteins: Catalytic/Structural and Regulatory*, pp. 11–30. Edited by J. F. Kane. Boca Raton, FL. CRC Press.

Saier, M. H., Jr & Reizer, J. (1992). Proposed uniform nomenclature for the proteins and protein domains of the bacterial phosphoenolpyruvate:sugar phosphotransferase system. *J Bacteriol* **174**, 1433–1438.

Saier, M. H., Jr & Reizer, J. (1994). The bacterial phosphotransferase system: new frontiers 30 years later. *Mol Microbiol* **13**, 755–764.

Schwelberger, H. G., Kohlwein, S. D. & Paltauf, F. (1989). Molecular cloning, primary structure and disruption of the structural gene of aldolase from *Saccharomyces cerevisiae*. *Eur J Biochem* **180**, 301–308.

Sharp, P. M. & Li, W. H. (1987). The codon adaptation index – a

measure of directional synonymous codon usage bias, and its potential application. *Nucleic Acids Res* **15**, 1281–1295.

Singer, M., Rossmiessl, P., Cali, B. M., Liebke, H. & Gross, C. A. (1991a). The *Escherichia coli* *ts8* mutation is an allele of *fda*, the gene encoding fructose-1,6-diphosphate aldolase. *J Bacteriol* **173**, 6242–6248.

Singer, M., Walter, W. A., Cali, B. M., Rouviere, P., Liebke, H. H., Gourse, R. L. & Gross, C. A. (1991b). Physiological effects of the fructose-1,6-diphosphate aldolase *ts8* mutation on stable RNA synthesis in *Escherichia coli*. *J Bacteriol* **173**, 6249–6257.

Smith, D. W. (1988). A complete, yet flexible, system for DNA/protein sequence analysis using VAX/VMS computers. *Comput Appl Biosci* **4**, 212.

Smith, M. W., Feng, D.-F. & Doolittle, R. F. (1992). Evolution by acquisition: the case for horizontal gene transfers. *Trends Biochem Sci* **17**, 489–493.

Stolz, B., Huber, M., Markovic-Housley, Z. & Erni, B. (1993). The mannose transporter of *Escherichia coli*. Structure and function of the IIAB^{Man} subunit. *J Biol Chem* **268**, 27094–27099.

Trach, K., Chapman, J. W., Piggot, P., LeCoq, D. & Hoch, J. A. (1988). Complete sequence and transcriptional analysis of the *spoOF* region of the *Bacillus subtilis* chromosome. *J Bacteriol* **170**, 4194–4208.

Valentin-Hansen, P., Hojrup, P. & Short, S. (1985). The primary structure of the DeoR repressor from *Escherichia coli* K-12. *Nucleic Acids Res* **13**, 5927–5936.

de Vos, W. M., Boerrigter, I. J., van Rooyen, R. J., Reiche, B. & Hengstenberg, W. (1990). Characterization of the lactose-specific enzymes of the phosphotransferase system in *Lactococcus lactis*. *J Biol Chem* **265**, 22554–22560.

Wehmeier, U. F. & Lengeler, J. W. (1994). Sequence of the *sor* operon for L-sorbose utilization from *Klebsiella pneumoniae* KAY2026. *Biochim Biophys Acta* **1208**, 348–351.

Wehmeier, U. F., Wöhr, B. M. & Lengeler, J. W. (1995). Molecular analysis of the phosphoenolpyruvate-dependent L-sorbose: phosphotransferase system from *Klebsiella pneumoniae* and of its multidomain structure. *Mol & Gen Genet* **246**, 610–618.

Wilson, R. and others (1994). 2.2 Mb of contiguous nucleotide sequence from chromosome III of *C. elegans*. *Nature* **368**, 32–38.

Yamada, M. & Saier, M. H., Jr (1988). Positive and negative regulators for glucitol (*gut*) operon expression in *Escherichia coli*. *J Mol Biol* **203**, 569–583.

Received 3 August 1995; accepted 20 September 1995.