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

We have analysed a gene cluster in the 674-769 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 IIAWn-like protein encoded in the 2-01 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encoding proteins of the phosphoenolpyruvate:sugar phosphotransferase system (PTS) (Reizer et al., 1994a, b, 1995; unpublished observations).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.

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 *Batella subtilis* (Martin-Verstraete et al., 1990) and the sorbose Enzyme II complex of *Klebsiella pneumoniae* (Wehmeier & Lengerter, 1994; Wehmeier et al., 1995) were homologous to the *E. coli* mannose system (Reizer et al., 1991b; Lengerter 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, two phosphorylation sites of the mannose permease are two histidine residues, whereas a histidine and a cysteine are phosphorylated in other PTS permeases (Erni, 1991; Martin-Verstraete et al., 1990; Lengerter, 1994), consisted of four domains IIA, IIBMan homologues; (4) the 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 IIBMan homologues); (4) *agaC* and *agaW* (two IICMan homologues); (5) *agaD* (IIDMan 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 (*pgG*) of *E. coli*, the twin *sac* genes (*sacP* and *sacX*) of *B. subtilis*, the *scr* operon (*scr-A*) 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<sup>Acb</sup> is found in the 24-41 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.

**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<sub>Man</sub> homologues); (4) *agaC* and *agaW* (two IIC<sub>Man</sub> homologues); (5) *agaD* (IID<sub>Man</sub> 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.

**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 DNAPEX system (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.

**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
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 aspartate_320 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 lysine_72 is conserved in AgaR, and lysine_80, lysine_13 and aspartate_210 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 RX_3[LIVMA]X_8[LIVM]X_20[T[LIVMA]R[KAN]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 0158 and ORF 0157, respectively)**

Two genes in the *aga* gene cluster, *agaB* and *agaV*, encode IIB\[^{Ma}\] 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* manno 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\[^{Fra}\](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 0169 (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 RDP2 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(Ata), 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 (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 egl 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).

<table>
<thead>
<tr>
<th>Protein</th>
<th>Accession no.</th>
<th>Identity (%)</th>
<th>Segments compared (no. of residues)</th>
<th>Comparison score (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GutR(Eco)</td>
<td>P15082</td>
<td>35</td>
<td>254</td>
<td>51</td>
</tr>
<tr>
<td>AccR(Ata)</td>
<td>P32104</td>
<td>30</td>
<td>239</td>
<td>37</td>
</tr>
<tr>
<td>GatR(Eco)</td>
<td>P36930</td>
<td>39</td>
<td>183</td>
<td>36</td>
</tr>
<tr>
<td>GipR(Eco)</td>
<td>P20932</td>
<td>28</td>
<td>251</td>
<td>34</td>
</tr>
<tr>
<td>ORF o269(Eco)</td>
<td>P32144</td>
<td>27</td>
<td>252</td>
<td>29</td>
</tr>
<tr>
<td>ORF232(Mca)</td>
<td>S48604</td>
<td>26</td>
<td>227</td>
<td>28</td>
</tr>
<tr>
<td>LacR(Lla)</td>
<td>P18816</td>
<td>28</td>
<td>244</td>
<td>28</td>
</tr>
<tr>
<td>DeoR(Eco)</td>
<td>P20617</td>
<td>27</td>
<td>230</td>
<td>28</td>
</tr>
<tr>
<td>SgcR(Eco)</td>
<td>P39361</td>
<td>26</td>
<td>251</td>
<td>27</td>
</tr>
<tr>
<td>FucR(Eco)</td>
<td>P11554</td>
<td>26</td>
<td>239</td>
<td>24</td>
</tr>
<tr>
<td>LacR(Smu)</td>
<td>P16644</td>
<td>25</td>
<td>232</td>
<td>23</td>
</tr>
<tr>
<td>ORF f251(Eco)</td>
<td>P39299</td>
<td>27</td>
<td>253</td>
<td>22</td>
</tr>
</tbody>
</table>

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<sup>Aga</sup> enzymes with the mannose-, sorbose- and fructose-specific Enzymes IIB

<table>
<thead>
<tr>
<th>Enzyme IIB</th>
<th>Identity (%)</th>
<th>Segments compared (no. of residues)</th>
<th>Comparison score (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IIB&lt;sup&gt;Aga&lt;/sup&gt; (157)</td>
<td>NA</td>
<td>IIB&lt;sup&gt;Aga&lt;/sup&gt; (158)</td>
<td>IIB&lt;sup&gt;Aga&lt;/sup&gt; (157)</td>
</tr>
<tr>
<td>IIB&lt;sup&gt;Aga&lt;/sup&gt; (158)</td>
<td>46</td>
<td>NA</td>
<td>153</td>
</tr>
<tr>
<td>IIB&lt;sup&gt;Bga&lt;/sup&gt; (168)</td>
<td>36</td>
<td>34</td>
<td>155</td>
</tr>
<tr>
<td>IIB&lt;sup&gt;Bga&lt;/sup&gt; (164)</td>
<td>35</td>
<td>30</td>
<td>153</td>
</tr>
<tr>
<td>IIB&lt;sup&gt;Bfr&lt;/sup&gt; (163)</td>
<td>32</td>
<td>27</td>
<td>158</td>
</tr>
</tbody>
</table>

NA, Not applicable.

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<sup>Mna</sup>, mannose-specific IIB protein domain of E. coli [IIB<sup>Mna</sup>(Eco), i.e. the C-terminal portion, previously designated the P20 protein domain of Enzyme IIB<sup>Mna</sup>; Erni et al., 1987, 1989]; IIB<sup>Sor</sup>, sorbose-specific Enzyme IIB of K. pneumoniae [IIB<sup>Sor</sup>(Kpn), Wehmeier & Lengeler, 1994; Wehmeier et al., 1995]; IIB<sup>FrU</sup>, fructose-specific Enzyme IIB of B. subtilis [IIB<sup>FrU</sup> (Bsu), Martin-Verstraete et al., 1990].

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 phosphorlylatable histidyl residue of IIB<sup>Mna</sup> (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], HGQ[LIVM]X,W, 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<sup>Mna</sup> family. The protein product of agaC (IICAga<sup>C</sup>) 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, IICAga<sup>B</sup>, 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) indicate that IICAga<sup>B</sup> is most similar to IICAga<sup>C</sup>, and that all five proteins are homologous.

Fig. 4(a) shows the multiple alignment of the five homologous members of the IIC<sup>Mna</sup> 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 IICAga<sup>C</sup> 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<sup>Mna</sup> 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<sup>Aga</sup> (see below).

Within the N-terminal regions of the IIC<sup>Mna</sup> family proteins is a cysteinyl residue (alignment position 36) which is conserved in IIC<sup>Mna</sup>, IIC<sup>FrU</sup>, IIC<sup>Sor</sup> and IICAga<sup>C</sup> but not in the shorter but otherwise homologous IICAga<sup>B</sup> (Fig. 4a). Recent data have demonstrated that substitution of this cysteinyl residue in IIC<sup>Mna</sup> (C36S) does not affect phosphotransferase activity indicating that it is not an essential catalytic residue (Rhiel et al., 1994).
Fig. 3. Multiple alignment (a) and mean similarity plot (b) of the proteins comprising the IIBMan 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 IIBMan 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 IICMan 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 IICMan-like proteins the N-terminal sequence is localized...
Table 3. Comparison of the amino acid sequences of the two IIC\textsuperscript{Ag} proteins with the mannose-, sorbose- and fructose-specific Enzyme IIC proteins

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>Identity (%)</th>
<th>Segments compared (no. of residues)</th>
<th>Comparison score (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IIC\textsuperscript{Ag}\textsuperscript{A} (133)</td>
<td>IIC\textsuperscript{Ag}\textsuperscript{A} (267)</td>
<td>IIC\textsuperscript{Ag}\textsuperscript{A} (133)</td>
</tr>
<tr>
<td>IIC\textsuperscript{Ag}\textsuperscript{A} (133)</td>
<td>NA</td>
<td>39</td>
<td>NA</td>
</tr>
<tr>
<td>IIC\textsuperscript{Ag}\textsuperscript{A} (267)</td>
<td>39</td>
<td>NA</td>
<td>124</td>
</tr>
<tr>
<td>IIC\textsuperscript{Ag}\textsuperscript{A} (Eco) (266)</td>
<td>36</td>
<td>26</td>
<td>108</td>
</tr>
<tr>
<td>IIC\textsuperscript{Ag}\textsuperscript{A} (Kpn) (266)</td>
<td>36</td>
<td>27</td>
<td>112</td>
</tr>
<tr>
<td>IIC\textsuperscript{Ag}\textsuperscript{A} (Bu) (269)</td>
<td>38</td>
<td>24</td>
<td>123</td>
</tr>
</tbody>
</table>

NA, Not applicable.

Table 3. Comparison of the amino acid sequences of the two IIC\textsuperscript{Ag} 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\textsuperscript{Mn}, mannose-specific IIC protein of E. coli [IIC\textsuperscript{Mn}(Eco), previously named II-P\textsuperscript{Man}, Erni et al., 1987]; IIC\textsuperscript{So}, sorbose-specific IIC protein of K. pneumoniae [IIC\textsuperscript{So}(Kpn), Wehmeier & Lengeler, 1994; Wehmeier et al., 1995]; IIC\textsuperscript{FrU}, fructose-specific IIC protein of B. subtilis [IIC\textsuperscript{FrU}(Bsu), Martin-Verstraete et al., 1990].

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>Identity (%)</th>
<th>Segments compared (no. of residues)</th>
<th>Comparison score (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IIC\textsuperscript{Ag}\textsuperscript{A} (133)</td>
<td>IIC\textsuperscript{Ag}\textsuperscript{A} (267)</td>
<td>IIC\textsuperscript{Ag}\textsuperscript{A} (133)</td>
</tr>
<tr>
<td>IIC\textsuperscript{Ag}\textsuperscript{A} (133)</td>
<td>NA</td>
<td>39</td>
<td>NA</td>
</tr>
<tr>
<td>IIC\textsuperscript{Ag}\textsuperscript{A} (267)</td>
<td>39</td>
<td>NA</td>
<td>124</td>
</tr>
<tr>
<td>IIC\textsuperscript{Ag}\textsuperscript{A} (Eco) (266)</td>
<td>36</td>
<td>26</td>
<td>108</td>
</tr>
<tr>
<td>IIC\textsuperscript{Ag}\textsuperscript{A} (Kpn) (266)</td>
<td>36</td>
<td>27</td>
<td>112</td>
</tr>
<tr>
<td>IIC\textsuperscript{Ag}\textsuperscript{A} (Bu) (269)</td>
<td>38</td>
<td>24</td>
<td>123</td>
</tr>
</tbody>
</table>

NA, Not applicable.

Fig. 5 presents the multiple alignment (a), the mean hydropathy plot (b) and the mean similarity plot (c) of this IID family. Approximately 20% of the residues are fully conserved in the four members of the IID\textsuperscript{Man} family. The higher similarity among proteins of the IID\textsuperscript{Man} or IIC\textsuperscript{Man} families, as compared to the similarity among members of the IIB\textsuperscript{Man} family is documented by the mean similarity values calculated for the entire alignment of the IID\textsuperscript{Man} homologues - 0.85 (dashed line in Fig. 5c), the IIC\textsuperscript{Man} homologues - 0.83 (dashed line in Fig. 4c), and the IIB\textsuperscript{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 IID\textsuperscript{Man}, IID\textsuperscript{So} and IIC\textsuperscript{FrU} (Lengeler et al., 1994; Wehmeier et al., 1995) is not
conserved in the homologous IID\textsuperscript{Asc} 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\textsuperscript{Asc}. A recent report has shown that this cysteinyl is not essential for PTS activity since the C43E mutant of IID\textsuperscript{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\textsuperscript{Man} family. Convention of presentation is as described in the legend to Fig. 3. The signature sequence of this IID\textsuperscript{Man} protein family is provided above the mean similarity plot. Abbreviations and references to the published sequences are as follows: IID\textsuperscript{Aga}(Eco), IID\textsuperscript{Aga} protein of \textit{E. coli} (GenBank accession no. U18997); IID\textsuperscript{FrU}(Bsu), fructose-specific IID protein of \textit{B. subtilis} (Martin-Verstraete et al., 1990); IID\textsuperscript{Man}(Eco), mannose-specific IID protein of \textit{E. coli} (previously named II-M\textsuperscript{Man}, Erni et al., 1987); IID\textsuperscript{Sor}(Kpn), sorbose-specific IID protein of \textit{K. pneumoniae} (Wehmeier & Lengeler, 1994; Wehmeier et al., 1995).

**AgA (ORF104)**

The \textit{aga} operon does not encode a II\textsuperscript{Man} homologue. However, screening the current databases (GenBank version 89.0; Swiss-Prot version 31.0) revealed an ORF (GenBank accession no. D26562; Fujita \textit{et al.}, 1994) in the 2.4-4.1 min region of the \textit{E. coli} chromosome which is homologous to II\textsuperscript{Man}. Comparison scores of this ORF with II\textsuperscript{Man}, II\textsuperscript{FrU} and II\textsuperscript{Sor} were 10-13 SD (23-35\% identity), thereby establishing that these four II\textsuperscript{A} proteins are homologous. II\textsuperscript{Aga} is 104 residues long whereas II\textsuperscript{Man} (136 residues), II\textsuperscript{Sor} (135 residues) and II\textsuperscript{FrU} (146 residues), are all substantially longer. The II\textsuperscript{Aga} protein proved to lack the N-terminal 40 residue segment which includes the reported active site histidyl phosphorylation site (His\textsubscript{104}) in II\textsuperscript{Man} (Erni \textit{et al.}, 1989). Although residue 11 in II\textsuperscript{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.
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$^\text{Man}$, Enzyme II$^\text{Sor}$, Enzyme II$^\text{Fru}$, and Enzyme II$^\text{AgA}$ 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$^\text{AgA}$ and IIC$^\text{AgA}$) these partial sequences exhibit long branch lengths as expected. Only in the case of the IIB family were the twin II$^\text{AgA}$ 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 0163)

The aga A 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 nag A gene
Table 4. Binary comparisons of protein members of the glucosamine-6-phosphate-deaminase isomerase family

<table>
<thead>
<tr>
<th></th>
<th>NagB(Eco) (266)</th>
<th>Nag1(Cal) (248)</th>
<th>ORF213(Eco) (213)</th>
<th>ORF289(Hsa) (289)</th>
<th>ORF126(Mca) (126) fragment</th>
</tr>
</thead>
<tbody>
<tr>
<td>AgaI(Eco)</td>
<td>28 (201)</td>
<td>25 (218)</td>
<td>26 (203)</td>
<td>27 (206)</td>
<td>24 (108)</td>
</tr>
<tr>
<td>(251)</td>
<td>26</td>
<td>24</td>
<td>16</td>
<td>30</td>
<td>11</td>
</tr>
<tr>
<td>NagB(Eco)</td>
<td>47 (252)</td>
<td>27 (212)</td>
<td>60 (258)</td>
<td>34 (126)</td>
<td></td>
</tr>
<tr>
<td>(266)</td>
<td>76</td>
<td>18</td>
<td>101</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>Nag1(Cal)</td>
<td>26 (220)</td>
<td>47 (252)</td>
<td>33 (126)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(248)</td>
<td>15</td>
<td>62</td>
<td>17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ORF213(Eco)</td>
<td></td>
<td>28 (207)</td>
<td>24 (90)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(213)</td>
<td></td>
<td>17</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ORF289(Hsa)</td>
<td></td>
<td></td>
<td></td>
<td>35 (126)</td>
<td></td>
</tr>
<tr>
<td>(289)</td>
<td></td>
<td></td>
<td></td>
<td>23</td>
<td></td>
</tr>
</tbody>
</table>

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 (ktrp = 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 average 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 Candida albicans [Nag1(Cal), Natarajan & Datta, 1993]; ORF213, hypothetical 23.5-kDa protein encoded within the trnB-bglB (81-5-84-5 min) region of the E. coli chromosome [ORF213(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].

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 cysteiny1 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 cysteiny1s 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, Candid. 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...
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. stearothermophilus and the E. coli agaY gene product. 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
N-Acetylgalactosamine-utilization gene cluster

**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 dipptide 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_ (Alefoudou _et al._, 1989; Alefoudou & Perham, 1989), Coryne. glutamicum (von der Osten _et al._, 1989) and _B. subtilis_ (putative) (Trach _et al._, 1988; Mitchell _et al._, 1992), respectively; GardY(Eco) and AgaY(Eco), putative tagatose-1,6-sisphosphate 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 (SBAs) of _R. sphaeroides_ (SBAI (Rsp), previously named CkB, Gibson _et al._, 1991; and SBAII (Rsp), previously named CFB, Chen _et al._, 1991). Thick lines segregate binary comparisons of (1) the characterized FBPs 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).

<table>
<thead>
<tr>
<th></th>
<th>FBA(Spo) (358)</th>
<th>FBA(Eco) (359)</th>
<th>FBA(Cgl) (344)</th>
<th>SBAII(Rsp) (354)</th>
<th>SBAI(Rsp) (359)</th>
<th>FBA(Bsu) (285)</th>
<th>AgaY(Eco) (286)</th>
<th>GardY(Eco) (286)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FBA(Spo) (359)</td>
<td>67 (360)</td>
<td>49 (352)</td>
<td>37 (347)</td>
<td>33 (51)</td>
<td>40 (43)</td>
<td>33 (54)</td>
<td>34 (44)</td>
<td>29 (133)</td>
</tr>
<tr>
<td>FBA(Spo) (358)</td>
<td>164</td>
<td>127</td>
<td>60</td>
<td>11</td>
<td>6</td>
<td>7</td>
<td>5</td>
<td>11</td>
</tr>
<tr>
<td>FBA(Eco) (359)</td>
<td>50 (357)</td>
<td>37 (339)</td>
<td>35 (49)</td>
<td>36 (33)</td>
<td>31 (55)</td>
<td>41 (44)</td>
<td>30 (111)</td>
<td></td>
</tr>
<tr>
<td>FBA(Eco) (358)</td>
<td>122</td>
<td>77</td>
<td>19 (243)</td>
<td>15 (138)</td>
<td>28 (193)</td>
<td>34 (50)</td>
<td>28 (165)</td>
<td></td>
</tr>
<tr>
<td>FBA(Cgl) (359)</td>
<td>40 (351)</td>
<td>17 (321)</td>
<td>16 (339)</td>
<td>28 (167)</td>
<td>25 (154)</td>
<td>28 (47)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FBA(Cgl) (344)</td>
<td>63</td>
<td>7</td>
<td>4</td>
<td>11</td>
<td>13</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SBAII(Rsp) (354)</td>
<td>78 (356)</td>
<td>25 (123)</td>
<td>21 (135)</td>
<td>34 (152)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SBAI(Rsp) (359)</td>
<td>116</td>
<td>14</td>
<td>12</td>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SBAI(Rsp) (285)</td>
<td>18</td>
<td>40 (156)</td>
<td>31 (247)</td>
<td>35 (203)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FBA(Bsu) (286)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GardY(Eco) (286)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 coordination. Based on these observations and the multiple alignment shown in Fig. 8, we propose that the remaining histidyl, glutamyl and/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 thermostability 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]GX,G[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.
region III in the multiple alignment and contains three conserved acidic residues. The second signature sequence, [LIVM]AX[STAG]XGX,HGX,Y, 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 I1 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 0384)**

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 *RSEARCH* 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 *Myco. leprae* (Swiss-Prot accession no. P40831; 19% in 302 residues overlap; 10 SD), and of *Rhizobium milioli* 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 eu-ryotic 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 bears 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*-acetylglactosamine 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*-acetylglactosamine (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...
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 IIMan constituents), step 2 is catalysed by AgaA (a truncated version of NagA), step 3 is 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 Gat2 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 agaY 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 GlmS, 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-acetylglucosamine, 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-acetylglucosamine 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\textsuperscript{AgA} complex (IIA, IIB, IIC, IID and IIE) proved to be homologous to the previously established 'splinter' group of Enzymes II of the PTS (Reizer et al., 1991b), which includes the mannoside Enzyme II of E. coli, the fructose Enzyme II of B. subtilis and the recently sequenced sorbose Enzyme II of K. pneumoniae (Nobelm\textsuperscript{a}n & 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\textsuperscript{Mn}. 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\textsuperscript{Mn} can bypass (albeit at a slow rate) the IIA\textsuperscript{Mn} protein domain as an obligatory phosphoryltransfer intermediate (Stolz et al., 1993). It is possible either that IIA\textsuperscript{Mn} phosphorylates IIB\textsuperscript{AgA} and/or IIB\textsuperscript{AgB} or that the IIA function is not required. The truncated IIA\textsuperscript{AgA} identified here may be a non-functional remnant of a functional IIA\textsuperscript{AgA}. In this regard, it is relevant to note that IIA\textsuperscript{Mn} of the major PTS permease cluster phosphorylates IIB\textsuperscript{Gle}, IIB\textsuperscript{Ber}, IIB\textsuperscript{Bre} and IIB\textsuperscript{Mai} (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 hydrophathy 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 (Nobelmann & 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 *aga*Y 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 *aga*Y. 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*-acetylglucosamine 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*-acetylglucosamine.

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 *hh8*) (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*-acetylglucosamine. 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*AB* 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*AB* and IIB*AB* proteins may together form a heterooligomeric complex, and the same may be true of the IIC*AB* and IIC*AB* 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 5RO1AI21702 and 2RO1AI14176 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\textsubscript{Man}, IIA\textsubscript{Sor} and IIA\textsubscript{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


within the form |CO₂ fixation operon of Rhodobacter spheroides. J Biol Chem 266, 14646–14653.


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 63–68.


Reizer, J., Reizer, A. & Saier, M. H., Jr (1995). Novel phosphot-
transferase 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.


Received 3 August 1995; accepted 20 September 1995.