Analysis of the *Escherichia coli* genome. V. DNA sequence of the region from 76.0 to 81.5 minutes

Heidi J.Sofia, Valerie Burland, Donna L.Daniels, Guy Plunkett III and Frederick R.Blattner*
Laboratory of Genetics, University of Wisconsin, 445 Henry Mall, Madison, WI 53706, USA

Received March 23, 1994; Revised and Accepted May 23, 1994

GenBank accession no. U00039

ABSTRACT

The DNA sequence of a 225.4 kilobase segment of the *Escherichia coli* K-12 genome is described here, from 76.0 to 81.5 minutes on the genetic map. This brings the total of contiguous sequence from the *E.coli* genome project to 725.1 kb (76.0 to 92.8 minutes). We found 191 putative coding genes (ORFs) of which 72 genes were previously known, and 110 of which remain unidentified despite literature and similarity searches. Seven new genes—*arsE*, *arsF*, *arsG*, *treF*, *xylR*, *xylG*, and *xylH*—were identified as well as the previously mapped *pit* and *dctA* genes. The arrangement of proposed genes relative to possible promoters and terminators suggests 90 potential transcription units. Other features include 19 REP elements, 95 computer-predicted bends, 50 Chi sites, and one grey hole. Thirty-one putative signal peptides were found, including those of thirteen known membrane or periplasmic proteins. One tRNA gene (*proK*) and two insertion sequences (IS5 and IS150) are located in this segment. The genes in this region are organized with equal numbers oriented with or against replication.

INTRODUCTION

This is the fifth DNA segment sequenced by the *E.coli* Genome Project (1–4). The 225,419 basepairs (bp) described here brings the project up to 725,115 contiguous bp, slightly over 15.5% of the genome, with a total of 610 CDSs, 13 tRNA genes, and 12 rRNA genes. In this fifth segment, fewer of the putative ORFs had been previously recognized (one third compared with half the ORFs in our first four segments). *E.coli* is unique among the genomes currently being sequenced in the amount of existing biological data that can be incorporated into the analysis. We make a considerable effort to utilize this information to serve as a check and to guarantee the production of a definitive sequence. We have continued to analyze and annotate the sequence as accurately as possible and to develop and improve the methods to do this.

MATERIALS AND METHODS

The methods used in this project have been described in previous publications (1–4). Results from protein similarity searches are

reported as % scores from protein alignments. This score, or 'similarity index', represents % identity with a correction for gaps, and is named to reflect the fact that similarity as well as identity is considered in the initial alignment.

Some of the software used in the analysis of the sequence has been slightly modified to improve performance. Additional codon usage matrices have been incorporated into GenePlot (a graphical annotation tool) so that comparisons may be made between *E.coli* usage and that of other organisms. The matrices and algorithms were provided by M.Borodovsky (5).

Naming conventions

We use the term ORF as the standard short-hand for Open Reading Frame, but, in addition, use a naming convention that indicates the orientation and size of the ORF. Putative genes in the forward, or clockwise, direction are designated '*orf*' (or '*o*'), while those in the opposite direction are designated '*fro*' (or '*f*'). The number following the '*o*' or '*f*' indicates the length, in amino acids, of the predicted gene product. If possible we determine the amino terminus of the protein from the literature or from the Link database of experimentally determined N-terminal peptides (A.Link, personal communication). In the absence of other data we annotate the longest possible reading frame using an ATG or GTG start.

RESULTS AND DISCUSSION

The 225,419 bp sequence presented here has been deposited in the sequence databases as ECOUW76 and assigned the GenBank accession number U00039. This sequence lies to the left (counter clockwise) of the first four segments of our project and overlaps the adjacent segment ECOUW82 (GenBank accession L10328) by 547 bases. Figure 1 is a map of the sequence, illustrating the identified features. Genes and putative genes (ORFs), along with the predicted molecular weights and isoelectric points of the protein products, are listed in Table 1. Gene names are from published sequences or the last edition of the *E.coli* genetic map (6), unless otherwise noted. The entire sequence presented was newly determined from M13 subclone libraries of lambda encyclopedia clones as described (1). Previously published sequence was used only for comparison.

We have identified 191 putative genes (ORFs), including the partial gene *ggt* at the left end. Of these ORFs, 110 are potential

*To whom correspondence should be addressed

Table 1. Genes and predicted characteristics of deduced protein products

Gene ^a	Site No. ^b	Endpoints in sequence	first...last codon	molecular size (kD)	size (aa)	pI
<i>ggt^c</i>	18280	1085 < 1	ATG...TAA	61.8	580	5.4
<i>o146a (yhhA)</i>		1205 > 1645	ATG...TAA	16.6	146	10.9
<i>ugpQ</i>		2375 < 1632	ATG...TAG	27.4	247	6.3
<i>ugpC</i>	17794	3481 < 2372	GTG...TGA	41.0	369	8.4
<i>ugpE</i>	17791	4289 < 3444	ATG...TAA	31.5	281	6.1
<i>ugpA</i>	40	5173 < 4286	ATG...TGA	33.3	295	8.8
<i>ugpB</i>	39	6587 < 5271	ATG...TAA	48.5	438	6.5
<i>livF</i>		7711 < 6986	GTG...TAA	26.9	241	6.4
<i>livG</i>	553	8468 < 7701	ATG...TAA	28.4	255	7.0
<i>livM</i>	18190	9742 < 8465	ATG...TGA	46.3	425	9.7
<i>livH</i>	552	10665 < 9739	ATG...TGA	33.0	308	7.9
<i>livK</i>	550	11822 < 10713	ATG...TGA	39.4	369	5.5
<i>o127</i>		12246 > 12629	ATG...TAA	14.5	127	7.0
<i>livJ</i>	551	13977 < 12817	GTG...TAA	41.3	386	6.9
<i>rpoH (htrR, hin)</i>	618	15045 < 14191	ATG...TAA	32.5	284	5.7
<i>ftsX</i>	18298	16348 < 15290	ATG...TAA	38.5	352	9.0
<i>ftsE</i>	736	17009 < 16341	ATG...TAA	24.4	222	9.4
<i>ftsY</i>	18295	18505 < 17012	ATG...TAA	54.5	497	4.4
<i>o198 (yhhF)</i>		18655 > 19251	ATG...TAA	21.7	198	6.2
<i>o130</i>		19118 > 19510	ATG...TAA	15.0	130	10.5
<i>f119</i>		19872 < 19513	ATG...TAA	13.5	119	10.1
<i>o208</i>		20013 > 20639	ATG...TGA	23.8	208	9.3
<i>o732</i>		20713 > 22911	ATG...TAA	76.8	732	5.8
<i>f81</i>		23258 < 23013	ATG...TGA	9.1	81	5.1
<i>o221</i>		23479 > 24144	ATG...TAA	25.3	221	9.8
<i>o203</i>		24163 > 24774	GTG...TAA	22.1	203	5.8
<i>f419</i>		26037 < 24778	GTG...TAA	43.8	419	9.8
<i>o376</i>		26046 > 27176	GTG...TGA	41.4	376	10.2
<i>o195</i>		27231 > 27818	ATG...TGA	21.8	195	5.6
<i>nikA</i>		27929 > 29503	ATG...TAA	58.7	524	5.9
<i>nikB</i>		29503 > 30447	ATG...TGA	35.3	314	9.9
<i>nikC</i>		30444 > 31277	GTG...TAA	30.4	277	7.7
<i>nikD</i>		31277 > 32041	ATG...TGA	26.8	254	6.2
<i>nikE</i>		32038 > 32844	ATG...TAA	29.7	268	10.1
<i>o133 (yhhG)</i>		32850 > 33251	ATG...TGA	15.1	133	6.0
<i>rhsB</i>		33454 > 37689	ATG...TAA	159.4	1411	6.5
<i>o122 (yhhH)</i>		37676 > 38044	ATG...TAA	14.0	122	8.3
<i>o378 (yhhI)</i>		38640 > 39776	ATG...TAA	43.2	378	6.9
<i>f375 (yhhJ)</i>		41068 < 39941	ATG...TAA	41.2	375	7.5
<i>f648</i>		43011 < 41065	GTG...TGA	71.9	648	5.9
<i>f231</i>		43748 < 43053	GTG...TAA	25.7	231	7.4
<i>f355</i>		44862 < 43795	ATG...TGA	38.8	355	7.4
<i>f540</i>		46850 < 45228	ATG...TGA	61.2	540	6.3
<i>f123</i>		47483 < 47112	ATG...TAA	14.2	123	8.0
<i>f409</i>		48717 < 47488	ATG...TAA	47.0	409	6.9
<i>o383</i>		49001 > 50152	ATG...TAA	41.5	383	8.2
<i>f400</i>		51669 < 50467	GTG...TGA	43.8	400	6.1
<i>pit</i>	385	51901 > 53400	ATG...TAA	53.4	499	9.6
<i>f111</i>		53979 > 53644	ATG...TAA	13.0	111	8.9
<i>uspA</i>		54370 > 54804	ATG...TAA	16.1	144	5.0
<i>o489</i>		55121 > 56590	ATG...TAA	53.8	489	8.8
<i>f285</i>		57496 < 56639	GTG...TAA	31.0	285	8.4
<i>prlC</i>	18031	59441 < 57399	ATG...TGA	77.2	680	5.1
<i>o280a</i>		59644 > 60486	ATG...TAA	31.9	280	8.7
<i>gor</i>	665	60558 > 61910	ATG...TAA	48.8	450	5.8
<i>arsE</i>		62787 > 63140	ATG...TAA	13.3	117	8.0
<i>arsF</i>		63173 > 64483	GTG...TAA	46.3	436	8.6
<i>arsG</i>		64496 > 64921	ATG...TAA	15.9	141	5.4
<i>o260</i>		65550 > 66332	ATG...TGA	29.5	260	4.9
<i>f338 (ISS ORF)</i>		67457 < 66441	ATG...TAA	39.3	338	9.6
<i>slp</i>		68187 > 68786	ATG...TAA	22.2	199	6.6
<i>o176 (yhiF)</i>		68942 > 69472	ATG...TGA	20.4	176	8.7
<i>f215 (yhiD)</i>		70161 < 69514	ATG...TAA	23.2	215	7.0
<i>f112 (10K-S)</i>		70563 > 70225	ATG...TAA	12.5	112	7.0
<i>f110 (10K-L)</i>		70999 > 70667	ATG...TAA	11.9	110	4.9
<i>o190</i>		71254 > 71826	ATG...TAA	20.9	190	9.5
<i>o175</i>		72625 > 73152	ATG...TAG	20.6	175	9.7
<i>o385</i>		73491 > 74648	ATG...TAA	41.2	385	5.7
<i>o1037</i>		74673 > 77786	ATG...TAA	111.5	1037	5.3
<i>f242a</i>		78877 < 78149	ATG...TGA	28.0	242	8.6
<i>f274a</i>		80069 < 79245	ATG...TAG	31.6	274	8.9
<i>gadA</i>		81839 < 80439	ATG...TGA	52.7	466	5.2
<i>f465</i>		83447 < 82050	ATG...TAA	51.6	465	6.2
<i>trfF</i>		83851 > 85500	ATG...TAA	63.7	549	4.8
<i>f200</i>		86153 < 85551	ATG...TGA	22.6	200	6.6
<i>o323</i>		86601 > 87572	ATG...TGA	36.1	323	8.2
<i>o337</i>		87621 > 88634	ATG...TAA	37.9	337	9.8
<i>o440</i>		89045 > 90367	ATG...TAA	47.2	440	9.6
<i>f472</i>		91975 < 90557	GTG...TGA	50.5	472	8.8
<i>f183</i>		92622 < 92071	ATG...TAA	20.8	183	10.3
<i>f216</i>		93443 < 92793	ATG...TAG	25.3	216	7.3
<i>o382</i>		93456 > 94604	GTG...TAA	42.5	382	6.4
<i>f498</i>		96196 < 94700	ATG...TAA	55.5	498	5.8
<i>dclA</i>	872	97703 < 96417	ATG...TAA	45.4	428	8.8
<i>f651</i>		99841 < 97886	ATG...TAG	73.1	651	6.0
<i>f1165</i>		103453 < 99956	GTG...TAA	128.5	1165	6.1
<i>f368</i>		104513 < 103407	ATG...TAA	41.7	368	8.6
<i>f779</i>		106859 < 104520	ATG...TAA	86.0	779	5.7
<i>f692</i>		108948 < 106870	ATG...TGA	78.6	692	8.8
<i>f145</i>		109537 < 109100	GTG...TAA	17.3	145	11.2
<i>f242b</i>		110214 < 109486	GTG...TGA	27.1	242	5.3
<i>f62</i>		110438 < 110250	ATG...TAG	7.0	62	4.1
<i>o523</i>		110711 > 112282	ATG...TGA	59.4	523	6.4
<i>o62</i>		112282 > 112470	ATG...TGA	7.2	62	11.3
<i>o559</i>		112467 > 114146	ATG...TAA	62.0	559	6.4
<i>o423</i>		114816 < 116087	ATG...TAA	47.3	423	9.2
<i>dppF (dppE)</i>		117121 < 116117	ATG...TAA	37.6	334	8.9
<i>dppD</i>		118101 < 117118	ATG...TGA	35.8	327	6.4
<i>dppC</i>		119014 < 118112	ATG...TAA	32.3	300	6.6
<i>dppB</i>		120043 < 119024	ATG...TAA	37.5	339	8.7
<i>dppA</i>		121958 < 120351	ATG...TAA	60.3	535	6.5
<i>proK (proV)</i>	17632	122945 < 122869	(tRNA)			77 bp
<i>f589</i>		124760 < 122991	GTG...TAA	67.0	589	9.0
<i>f402</i>		126259 < 125051	ATG...TAA	43.0	402	9.2
<i>f234</i>		127192 < 126488	GTG...TAA	26.2	234	4.3
<i>kug</i>	124	127344 > 127907	ATG...TGA	21.1	187	6.5
<i>o146b</i>		127904 > 128344	ATG...TAA	17.1	146	6.7
<i>bisC</i>	951	130532 < 128313	ATG...TAA	81.9	739	5.7
<i>o219</i>		130799 > 131458	ATG...TAA	22.2	219	9.8
<i>o365</i>		131550 > 132647	ATG...TAA	40.1	365	6.7
<i>f276</i>		133415 < 132585	ATG...TAA	30.2	276	9.3
<i>o96</i>		133729 > 134019	ATG...TAG	11.0	96	8.1
<i>cspA</i>		134300 > 134512	ATG...TAA	7.4	70	5.7
<i>o173 (IS150 ORFA)</i>		134931 > 135452	ATG...TGA	19.7	173	10.1
<i>o283 (IS150 ORFB)</i>		135449 > 136300	GTG...TAA	33.5	283	9.7
<i>glyS (glySb)</i>	764	138648 < 136579	ATG...TAA	76.8	689	5.2
<i>glyQ (glySa)</i>		139569 < 138658	ATG...TAA	34.8	303	4.8
<i>o331</i>		140138 > 141133	ATG...TAA	37.6	331	9.5
<i>f146 (yiaA, URF4)</i>		141615 < 141175	GTG...TAA	16.3	146	4.9
<i>f117 (yiaB)</i>		142011 < 141658	ATG...TGA	13.0	117	8.7
<i>xylB</i>	4	143622 < 142168	ATG...TAA	52.6	484	5.6
<i>xylA</i>	5	145016 < 143694	ATG...TAA	49.7	440	6.0
<i>xylF</i>		145382 > 146374	ATG...TAA	35.7	330	5.4
<i>xylG</i>		146452 > 147993	ATG...TGA	56.5	513	6.0
<i>xylH</i>		147971 > 149152	ATG...TGA	41.0	393	10.4
<i>f274b</i>		149230 > 150408	ATG...TAG	44.9	392	6.4
<i>malS</i>	17659	151428 < 150604	ATG...TAA	30.9	274	10.5
<i>o417</i>		151748 > 153778	ATG...TAA	75.7	676	5.4
<i>o103</i>		153956 > 155209	ATG...TAA	46.7	417	5.8
<i>f282</i>		155384 > 155695	ATG...TAA	11.9	103	11.9
<i>o332</i>		156784 < 155936	ATG...TAA	31.1	282	6.7
<i>o155</i>		156985 > 157983	ATG...TGA	36.6	332	5.1
<i>o157a</i>		157995 > 158462	ATG...TAA	17.6	155	5.4
<i>o424</i>		158580 > 159053	ATG...TAA	17.5	157	8.3
<i>o328</i>		159056 > 160330	ATG...TAA	45.5	424	8.8
<i>o498</i>		160347 > 161333	ATG...TAA	36.0	328	7.0
<i>o220</i>		161337 > 162833	ATG...TGA	55.2	498	5.9
<i>o297</i>		162830 > 163492	ATG...TAA	23.4	220	5.3
<i>o231</i>		163452 > 164345	ATG...TAG	33.7	297	6.0
<i>f246</i>		164339 > 165034	ATG...TAA	25.6	231	5.8
<i>o324</i>		166121 < 165381	ATG...TAA	27.4	246	4.4
<i>f378a</i>		166245 > 167219	ATG...TAA	37.3	324	5.6
<i>f107</i>		168352 < 167216	ATG...TAA	41.8	378	9.1
<i>f542</i>		168681 < 168358	ATG...TAA	12.4	107	7.6
<i>f382</i>		170854 < 169226	GTG...TGA	59.7	542	5.7
<i>seIb (fdaA)</i>	17890	172077 < 170929	ATG...TAA	40.2	382	5.2
<i>seIa (fdaA)</i>	785	174113 < 172251	ATG...TAA	69.9	620	6.7
<i>f202 (yibF)</i>		175501 < 174110	ATG...TGA	50.6	463	6.3
<i>rhaA</i>		176207 < 175599	ATG...TGA	22.5	202	5.0
<i>o280b (yibA)</i>		176435 > 180568	ATG...TAG	156.3	1377	6.6
<i>o233 (yibJ)</i>		180589 > 181431	ATG...TAG	31.9	280	8.4
<i>o153 (yibG)</i>		181473 > 182174	GTG...TAA	26.6	233	4.9
<i>f378b (yibH)</i>		182429 > 182890	ATG...TAA	18.1	153	9.3
<i>f120 (yibI)</i>		185631 < 184495	ATG...TGA	42.0	378	7.6
<i>mlaA</i>	481	185996 < 185634	ATG...TAA	13.9	120	6.7
<i>mlaD</i>	479	186533				

new genes, still unidentified despite searches of the databases for similarities suggestive of function. Twenty-five of these ORFs had been sequenced previously but remain unidentified. Similarities of predicted ORFs with database protein sequences and matches to PROSITE patterns are summarized in Table 2.

Newly identified genes

Chromosomal arsenical resistance operon. The *o117*, *o436*, and *o141* translation products of the ORFs at 62787–64921 in the sequence are very similar to the plasmid-encoded arsenical resistance proteins *arsR*, *arsB*, and *arsC* produced by the *E. coli* R773 plasmid and similar resistance elements in staphylococcal species (See Table 2). Resistance is conferred by these plasmids through transport proteins that extrude toxic arsenic and antimony oxyanions (7). Arsenical resistance resembles other ATP-dependent transport systems in containing an ATP-binding protein and integral membrane components, but sequence similarity to other ATP-dependent transport proteins is limited.

The possibility of chromosomal arsenical resistance genes was suggested by the observation of partial resistance conferred by plasmids containing deletions in *ars* genes (7, 8). The assignment of *o117*, *o436*, and *o141* as an arsenical resistance locus is supported by previous results showing that arsenate sensitivity is tightly linked to the *gor* gene (9). A selection was performed that was designed to produce arsenate-resistant mutants by deleting the nearby *pit* gene, the low-affinity transporter of phosphate which is also the major transporter of arsenate, using thermal induction of a strain containing a Mu insertion in *gor*. Surprisingly, this approach resulted in strains that were more sensitive to arsenate, not less, than the *gor::Mu* strain, presumably because of deletions in an adjacent resistance determinant. This arsenate resistance locus was mapped to the 10 kb region immediately downstream of *gor* (9), which includes *o117*, *o436*, and *o141*. Data also exists that the product of *gor*, glutathione reductase, plays a role in arsenate resistance as well (10).

The *o117*, *o436*, and *o141* genes show much higher sequence similarity to the *E. coli* plasmid-encoded *ars* gene products than to the staphylococcal ones. However, the arrangement of the *E. coli* chromosomal genes matches that on the staphylococcal plasmids rather than the *E. coli* R773 plasmid. The *E. coli* plasmid operon has the structure *arsRDABC*, with *arsR* and *arsD* encoding regulators, *arsA* an inner membrane ATP-binding protein, *arsB* a membrane-spanning component, and *arsC* a subunit that determines substrate specificity. The arsenical resistance operons from staphylococcal species have the structure *arsRBC*. Like the *E. coli* chromosomal resistance locus, the staphylococcal system lacks the ATP-binding component, raising questions about the mechanism of energy transduction. Data suggests that these arsenical resistance systems are coupled to different energy sources depending upon subunit composition (11).

We suggest the names *arsE*, *arsF*, and *arsG* for the chromosomal arsenical resistance genes *o117*, *o436*, and *o141*.

Cytoplasmic trehalase. Trehalase is a highly conserved widely-distributed enzyme that splits the glucose disaccharide trehalose into glucose monomers. We suggest that *o549* at 83851–85500 encodes a cytoplasmic trehalase. The *o549* gene product contains two conserved motifs characteristic of trehalase: PS00927: TREHALASE_1 and PS00928: TREHALASE_2. The *o549* gene product is very similar to the previously identified periplasmic *E. coli* enzyme, with a similarity score of 50.3% over 495 aa,

but the similarity does not extend over the N terminus. The periplasmic TreA protein contains a typical signal sequence, but the *o549* gene product is very acidic in this region, indicating that the protein is not exported. Cytoplasmic extracts of strains that lack the periplasmic trehalase (*treA* mutants) have been observed to contain low levels of trehalase activity (12). This activity was independent of induction by trehalase but was repressed under conditions of low osmolarity. We suggest the name *treF* for *o549*.

***xylFGHR*.** D-Xylose is a component of plant hemicellulose and is the most abundant sugar in nature after glucose. *E. coli* can use D-xylose as a sole carbon and energy source by means of an inducible metabolic pathway consisting of transport, isomerase, and kinase activities that lead into the pentose phosphate pathway. *xylA* (145016–143694) encodes D-xylose isomerase and *xylB* immediately downstream encodes xylulose kinase. We assign the four ORFs that are divergently transcribed from *xylAB* to be high affinity transport genes and a regulatory gene for the xylose system. *o330* (145382–146374) is the same as the *xylF* gene mapped to the region, which encodes a gene necessary for binding-protein-dependent high affinity uptake of D-xylose. The D-xylose-binding protein was purified and the gene cloned; the sequence of the *xylF* gene was determined and published as an amino acid sequence, but it was not entered into the databases (13). Downstream of *xylF* are *o513*, *o393*, and *o392* which encode proteins that are similar to the ATP-binding and membrane-spanning components of ABC-type transporter systems and an AraC-type regulator respectively. Both this putative D-xylose transport system and the arabinose one contain a single membrane-spanning protein, which may be a homodimer, instead of two membrane-spanning proteins as in most periplasmic binding-protein-dependent transport systems. The gene designation *xylF* was suggested by analogy to the arabinose transport system, and similarly, we propose *xylG* as a name for the ATP-binding protein and *xylH* for the membrane component of the transporter. The gene designation *xylR* is suggested for the regulator. Pleiotropic *xyl⁻* phenotypes have been mapped 1–2 kb upstream of *xylA* (14, 15), suggesting that *xylR* may be the positive activator protein that coordinately regulates D-xylose high affinity transport and utilization (16). Immediately downstream of *xylAB* are the *f117* and *f146* genes, which some evidence suggests may also play some role in transport or regulation (14).

Identification of mapped genes

***pit*.** When not under conditions of phosphate limitation, *E. coli* mediates uptake of inorganic phosphate by the constitutive low-affinity transport system, Pit, which is also the main transporter of arsenate (17). Uptake is coupled to the proton motive force, is abolished by uncouplers, and is thought to depend upon a single membrane component. The *pit* gene has been mapped to min 77 near *gor* and cloned as a 2.2 kb *SalI*–*AvaI* fragment (18). Comparison of the restriction map of this fragment with our sequence revealed that *pit* is *o499* (51901–53400). No significant similarities with the predicted product of *o499* can be detected with other protein sequences.

***dctA*.** We assign *dctA* to *f428* (97703–96417). Aerobic transport of C-4 dicarboxylates depends upon the function of at least three separate genes known as *dctA*, *dctB*, and *cbt* (19). *dctA* has been mapped to min 80 near *proK* and *glyS*, and encodes a membrane-

spanning protein with a dicarboxylate-binding site on the periplasmic face (20). The *f428* gene in this region encodes a putative protein with a conserved sodium:dicarboxylate symporter site and high similarity to C-4 dicarboxylate permease proteins, with similarity scores of 62.8% over 422 aa to DctA from *Rhizobium leguminosarum* and 61.7% over 418 aa to DctA from *Rhizobium meliloti*. The *dctA* gene is immediately adjacent to genes that are similar to cellulose synthesis proteins. Another gene in this segment, *o328* (160347–161333) encodes a protein similar to DctP, the periplasmic binding protein component of a distinct C-4 dicarboxylate transport system in *Rhodobacter capsulatus* (21). *o328* is adjacent to genes that may function in carbohydrate metabolism, as discussed below.

o385 and *o1037*. These ORFs may define the defect in the *E. coli* strain PM61, a cell-division mutant with pleiotropic properties that were attributed to the *envC* locus which was originally mapped to 81 min (22). Complementation of the *envC* phenotype was used to clone a locus which mapped to min 74, and the genes were sequenced and named *envC* and *envD* (23, 24). In addition, the *acrA* and *acrE* genes mapping at 10 minutes show similar properties (25) and their predicted products share a similarity score of 72–75% with EnvC and EnvD. The *acrAE* and *envCD* loci also have been sequenced and entered into the database as *acrAB* (26) and *acrEF* (27), respectively. We find a third similar locus at 73491–77786 consisting of *o385* and *o1037*. The putative *o385* product has a similarity score of 49% over 371 aa to EnvC and the putative *o1037* product has a score of 70% over 1032 aa to EnvD. Given the closeness of these genes to the map position of the original mutation, this new locus may represent the true defect in PM61.

kdgK. The degradation of 3-deoxy-2-oxo-D-gluconate (KDG) depends upon the KDG kinase encoded by the *kdgK* gene, which has been mapped between *gor* and *dctA* (6, 28). The best candidate for *kdgK* in this region is *o382* (93456–94604). The predicted product of *o382* is similar to fructokinase from several enteric species and has a close match to one of two conserved sites in the PfkB family of carbohydrate kinases (PFKB_KINASES_2) as well as a partial match to the other conserved site (PFKB_KINASES_1).

pcsA. The phenotype of *pcsA* mutants includes cold-sensitive inhibition of cell division and sensitivity to detergents. The *pcsA* locus was mapped to 82 minutes between *rfa* and *rpmGB*; however, all the genes in this region have now been identified (6). As previously suggested, some *rfaG* mutants have a similar phenotype, indicating that *pcsA* may be allelic to *rfaG* (221331–220207) (29).

Unidentified genes

Some ORF products show strong similarities to other proteins but the data is insufficient for a firm identification.

bcs. *E. coli* is not considered to be a cellulose-producing bacterium. However, three gene products, encoded by *f692*, *f779*, and *f1165* at 108948–99956 in this segment show similarity to the *bcs* operon cellulose synthesis proteins of *Acetobacter xylinum*, *BcsA*, *BcsB*, and *BcsC*. Although debate still exists about whether *BcsA* or *BcsB* is the cellulose synthase protein, all three proteins are required for cellulose production *in vivo* (30, 31, 32). *f692* encodes a protein with a similarity score of 44.7% over

Table 2. Similarity of predicted ORF products to other proteins. (newly identified genes and unidentified ORFs, presented in map order)

ORF	length	Sequence similarity and/or PROSITE match	score
<i>o146a</i>	146 aa	Similar to several glutamine- and asparagine-rich proteins: JQ1150 (PIR) 648 <i>Dicystostelium discoideum</i> protein kinase 2 A26892 (PIR) 139 Mouse Mopa box protein (fragment) AAC3_DICDI (SW) 437 <i>Dicystostelium discoideum</i> ORF product (fragment) DROMASTM_1 (GB) 1655 <i>Drosophila virilis</i> mastermind gene	25.2% 140 aa 25.4% 117 aa 29.9% 87 aa 30.1% 72 aa
<i>o198</i>	198 aa	PS00092: N6_MTASE [N-6 Adenine-specific DNA methylases signature]	100% match
<i>o732</i>	732 aa	PS00154: ATPASE_E1_E2 [E1-E2 ATPases phosphorylation site] PS00339: AA_TRNA_LIGASE_IL_2 [Aminoacyl-transfer RNA synthetases class-II signatures]	100% match 100% match
		Similar to several cation transport ATP-ase family proteins [E1_E2 ATP-ases]: CADA_BACFI (SW) 723 <i>Bacillus firmus</i> CdaA, cadmium efflux ATP-ase CADA_STAAU (SW) 727 <i>Staphylococcus aureus</i> CdaA, cadmium efflux ATP-ase ATKA_ENTFA (GB) 727 <i>Enterococcus faecalis</i> CopA, potassium/copper transporting ATP-ase FIXL_RHIME (SW) 757 <i>Rhizobium meliloti</i> FixL, symbiotic nitrogen fixation cation ATP-ase ATKR_ECOLI (SW) 682 <i>Escherichia coli</i> KdpB, potassium transporting ATP-ase YEDF_ECOLI (SW) 77 <i>Escherichia coli</i> ORF9 hyp. protein in <i>amyA-fliE</i> intergenic region STYAMYA_4 (GB) 77 <i>Salmonella typhimurium</i> ORF product, hypothetical 8.6 kD protein ANAKETPOL_3 (GB) 505 <i>Anabaena</i> species ORF product, polyketide synthesis	36.5% 708 aa 35.2% 718 aa 33.2% 612 aa 33.9% 533 aa 27.7% 528 aa 30.6% 72 aa 30.6% 72 aa 30.3% 106 aa 28.3% 95 aa
<i>o195</i>	195 aa	ECOUW82_76 (GB) 252 <i>Escherichia coli</i> o252 product	100% match
<i>f375</i>	375 aa	PS00890: ABC2_MEMBRANE [ABC-2 type transport system integral membranes proteins]	100% match
<i>f648</i>	648 aa	PS00017: ATP_GTP_A [ATP/GTP-binding site motif A (P-loop)] PS00211: ABC_TRANSPORTER [ABC transporters family signature] (matches 6 out of 8 sites in 12 aa pattern; has P instead of S at position 2 and L for [SAG] at position 3)	100% match close match
		Similar to several transporter proteins: NODI_BRAJA (SW) 306 <i>Bradyrhizobium japonicum</i> NodI, putative oligosaccharide export protein NODI_RHILV (SW) 311 <i>Rhizobium leguminosarum</i> NodI, putative oligosaccharide export protein NODI_RHILLO (SW) 339 <i>Rhizobium loti</i> NodI, putative oligosaccharide export protein DRRA_STRPE (SW) 330 <i>Streptomyces pectus</i> DrrA, daunorubicin resistance protein	35.7% 242 aa 35.6% 234 aa 34.0% 236 aa 35.5% 229 aa
<i>f231</i>	231 aa	PS00017: ATP_GTP_A [ATP/GTP-binding site motif A (P-loop)] PS00211: ABC_TRANSPORTER [ABC transporters family signature] (matches 7 out of 8 sites in 12 aa pattern; has C instead of [SAG] at position 12) STMOLC_4 (GB) 325 <i>Streptomyces antibioticus</i> ATP-binding oleandomycin export protein MALK_ECOLI (SW) 371 <i>Escherichia coli</i> MalK, ATP-binding maltose transport protein MALK_SALTY (SW) 369 <i>Salmonella typhimurium</i> MalK, ATP-binding maltose transport protein ABC_ECOLI (SW) 203 <i>Escherichia coli</i> ATP-binding protein in ABC transporter family	100% match close match 33.0% 222 aa 29.4% 218 aa 29.0% 218 aa 29.1% 180 aa
<i>f123</i>	123 aa	Similar to several mammalian D-amino acid oxidase proteins, for example: OXDA_HUMAN (SW) 347 Human D-amino acid oxidase	33.3% 51 aa
<i>f111</i>	111 aa	YSCDHICIA_1 (GB) 3457 <i>Saccharomyces cerevisiae</i> DHC1, dynein	34.8% 46 aa
<i>o280a</i>	280 aa	PS00092: N6_MTASE [N-6 Adenine-specific DNA methylases signature] YORJ_HAEIN (SW) 281 <i>Haemophilus influenzae</i> ORF, hypothetical 32.2 kD protein	100% match 66.0% 281 aa
<i>arsE</i>	117 aa	PS00846: HTH_ARSR_FAMILY [Bacterial regulatory proteins, Arsr family signature]	100% match
		Similar to several Arsr proteins: ARSR_ECOLI (SW) 117 <i>Escherichia coli</i> Arsr, transcriptional regulator ARSR_STAAU (SW) 104 <i>Staphylococcus aureus</i> Arsr, transcriptional activator ARSR_STAXY (SW) 104 <i>Staphylococcus xylosum</i> Arsr, transcriptional activator	77.0% 113 aa 41.0% 77 aa 35.6% 86 aa
<i>arsF</i>	436 aa	Similar to several Arsf proteins: ARSB_ECOLI (SW) 429 <i>Escherichia coli</i> Arsf, arsenical pump membrane protein ARSB_STAAU (SW) 429 <i>Staphylococcus aureus</i> arsenical pump membrane protein ARSB_STAXY (SW) 429 <i>Staphylococcus xylosum</i> arsenical pump membrane protein ARSC_ECOLI (SW) 141 <i>Escherichia coli</i> Arsf, arsenical pump modifier YGLI_STRVR (SW) 119 <i>Streptomyces viridochromogenes</i> ORF1 product in <i>glnI</i> region D27733 (PIR) 147 <i>Azotobacter vinelandii</i> ORF3 product in <i>nif</i> locus	90.7% 429 aa 59.2% 426 aa 59.2% 426 aa 94.3% 141 aa 30.0% 118 aa 25.8% 96 aa
<i>arsG</i>	141 aa	PS00622: HTH_LUXR_FAMILY [Bacterial regulatory proteins, LuxR family signature] (matches 7 out of 11 conserved or partially conserved sites in 28 residue pattern)	close match
<i>o176</i>	176 aa	Similar to several regulators in LuxR family in region containing conserved motif: FIMZ_SALTY (SW) 210 <i>Salmonella typhimurium</i> FimZ, putative regulator in LuxR family FIXJ_AZOCA (SW) 211 <i>Azotobacter caulinodans</i> FixJ, regulator of <i>nif</i> and <i>fix</i> genes FIXJ_RHIME (SW) 204 <i>Rhizobium meliloti</i> FixJ, regulator of <i>nif</i> and <i>fix</i> genes FIMZ_ECOLI (SW) 231 <i>Escherichia coli</i> FimZ, putative regulator in LuxR family FIXJ_BRAJA (SW) 205 <i>Bradyrhizobium japonicum</i> FixJ, regulator of <i>nif</i> and <i>fix</i> genes	36.5% 63 aa 33.3% 54 aa 33.3% 54 aa 28.8% 79 aa 25.0% 60 aa
<i>f215</i>	215 aa	ATMC_SALTY (SW) 231 <i>Salmonella typhimurium</i> MgtC, Mg ²⁺ transport ATP-ase	31.9% 180 aa
<i>o385</i>	385 aa	Similar to acriflavin resistance proteins: ECU00734_1 (GB) 397 <i>Escherichia coli</i> AcrA, acriflavin resistance protein ENVV_ECOLI (SW) 385 <i>Escherichia coli</i> EnvV, acriflavin resistance protein PS00017: ATP_GTP_A [ATP/GTP-binding site motif A (P-loop)]	54.8% 373 aa 49.1% 371 aa 100% match
<i>o1037</i>	1037 aa	Similar to acriflavin resistance proteins: ECU00734_2 (GB) 1049 <i>Escherichia coli</i> AcrE, acriflavin resistance protein ENVV_ECOLI (SW) 1034 <i>Escherichia coli</i> EnvD, acriflavin resistance protein YFFA_ECOLI (SW) 40 <i>Escherichia coli</i> YffA (fragment), hypothetical protein in <i>dapE</i> 5' region PS00041: HTH_ARAC_FAMILY [Bacterial regulatory proteins, AraC family signature]	71.0% 1035 aa 70.0% 1032 aa 70.6% 34 aa 100% match
<i>f242a</i>	242 aa	Similar to several regulators in AraC family controlling outer-membrane protein expression: APPY_ECOLI (SW) 243 <i>Escherichia coli</i> AppY, growth-phase dependent regulator ENVV_ECOLI (SW) 253 <i>Escherichia coli</i> EnvY, porin thermoregulatory protein RNS_ECOLI (SW) 265 <i>Escherichia coli</i> Rns, adhesin expression regulator A40644 (PIR) 293 <i>Proteus mirabilis</i> UreR, urease operon transcriptional regulator	36.0% 234 aa 39.8% 127 aa 27.8% 242 aa 27.8% 132 aa
<i>f274a</i>	274 aa	PS00041: HTH_ARAC_FAMILY [Bacterial regulatory proteins, AraC family signature] Similar to several regulators in AraC family controlling outer-membrane protein expression: ENVV_ECOLI (SW) 253 <i>Escherichia coli</i> EnvY, porin thermoregulatory protein STYFEBACD_13 (GB) 259 <i>Salmonella typhimurium</i> ORF11 product on 90 kb virulence plasmid APPY_ECOLI (SW) 243 <i>Escherichia coli</i> AppY, growth-phase dependent regulator ECM5_1 (GB) 249 <i>Escherichia coli</i> M5 polypeptide RNS_ECOLI (SW) 265 <i>Escherichia coli</i> Rns, adhesin expression regulator	40.1% 186 aa 40.4% 175 aa 38.3% 131 aa 26.1% 239 aa 23.9% 201 aa
<i>f465</i>	465 aa	PS00190: CYTOCHROME_C [Cytochrome c family heme-binding site signature] Three sites S05001 (PIR) 302 <i>Pseudomonas aeruginosa</i> cytochrome c peroxidase	100% match 39.2% 278 aa
<i>trf</i>	549 aa	PS00927: TREHALASE_1 [Trehalase signatures] PS00928: TREHALASE_2 [Trehalase signatures]	100% match 100% match
		Similar to several trehalase enzymes: TREA_ECOLI (SW) 565 <i>Escherichia coli</i> TreA, osmoregulated periplasmic trehalase TREA_RABIT (SW) 578 Rabbit TreA, intestinal trehalase TREA_BONMO (SW) 579 <i>Bomx mori</i> (silk moth) gut trehalase TREA_TENMO (SW) 555 <i>Tenebrio molitor</i> (yellow mealworm) male accessory gland trehalase	50.3% 495 aa 33.0% 525 aa 30.9% 528 aa 30.4% 524 aa
<i>f200</i>	200 aa	PS00622: HTH_LUXR_FAMILY [Bacterial regulatory proteins, LuxR family signature] (matches 10 out of 11 sites in 28 aa pattern; has R instead of K at position 28.) Similar to transcriptional regulators in LuxR family: NARL_ECOLI (SW) 216 <i>Escherichia coli</i> NarL, nitrate/nitrite response regulator DEGU_BACSU (SW) 229 <i>Bacillus subtilis</i> DegL, regulator of protease production	close match 23.5% 201 aa 42.4% 59 aa
<i>o323</i>	323 aa	PS00044: HTH_LYSR_FAMILY [Bacterial regulatory proteins, LysR family signature]	100% match
		Similar to several proteins in LysR family: YAPC_ECOLI (SW) 304 <i>Escherichia coli</i> YafC, putative transcriptional regulator JC1120 (PIR) 306 <i>Pseudomonas</i> sp. SdsB, transcriptional activator for SDS biodegradation YOPE_PSEAE (SW) 249 <i>Pseudomonas aeruginosa</i> ORF product, putative transcriptional regulator	28.8% 292 aa 27.2% 290 aa 30.6% 242 aa
<i>o440</i>	440 aa	PS00216: SUGAR_TRANSPORT_1 [Sugar transport proteins signature 1]	close match

659 aa and 48.8% over 613 aa to the BcsA protein from two different strains of *A. xylinum* (30, 31). *f779* encodes a protein with a score of 26.7% over 780 aa and 26.9% over 776 aa to the BcsB proteins from these two strains (30, 32). NodC has been compared with several polysaccharide synthesis proteins, including BcsA, by multiple sequence alignment; conserved features were observed that can also be seen in the *f692* product (33). *f1165* encodes a protein with a score of 23.7% over 692 aa to the BcsC protein. A fourth gene, *f368*, appears to be located in a possible transcription unit with the three *bcs* homologues and encodes a protein similar to several cellulases in the D family, including the cellulase upstream of the Bcs operon in *A. xylinum* (34). *E. coli* was used as a negative control in an experiment where various bacterial and plant species were probed with antibodies to the BcsB protein (35). Special growth conditions may be necessary to induce these *E. coli bcs* genes.

Alcohol dehydrogenase and acetaldehyde dehydrogenase. *Zymomonas mobilis* and *Saccharomyces cerevisiae* both produce ethanol and carbon dioxide as primary fermentation products. *f382* (172077–170929) shows very high similarity to the NAD-dependent iron-activated alcohol dehydrogenases found in these microorganisms, with a score of 63.8% over 378 aa to the ADH2 bacterial protein and 54.5% over 377 aa to the ADH4 yeast protein. The predicted product of *f382* also shows high similarity to other proteins in the type III alcohol dehydrogenase family, including methanol dehydrogenase from *Bacillus methanolicus*, propanediol oxidoreductase from *E.coli*, and the carboxy-terminal region of the *adhE* gene product of *E.coli*, a multifunctional protein with alcohol dehydrogenase, aldehyde dehydrogenase, and pyruvate-formate-lyase deactivase activities. Immediately downstream of *f382* is *f542* which has a similarity score of 70.6% over 500 aa to the aldehyde dehydrogenase II from *Alcaligenes eutrophus*, a protein which is proposed to catalyze the oxidation of acetaldehyde to acetate in both ethanol and acetoin metabolism (36). The apparent organization of *f382* and *f542* in a possible transcription unit is consistent with the activities of these proteins being linked in a metabolic pathway.

gluE. The predicted product of *o157b* at 195465–195938 has a similarity score of 40.4% over 154 residues to a *Bacillus stearothermophilus* ORF product of unknown function and 26.2% over 162 aa to SpoU of *E.coli* (37). Since *o157b* is the only unidentified gene between *lct* and *cysE* (6), it may be a candidate for the *gluE* locus which was mapped to this region. *gluE* is associated with temperature-sensitive changes in the glutamyl-

tRNA synthetase affinity for glutamate (38), and is distinct from *gluX* at 52 min, the structural gene for glutamyl-tRNA synthetase. *gluE* was suggested to encode a regulatory protein, possibly an unidentified co-purifying 46 kD polypeptide (39), which is also known to affect the substrate affinities of the synthetase (40). The predicted molecular weight of the *o157* product, however, is 17.7 kD. Other proteins that copurify with glutamyl-tRNA synthetase have been identified (41–43).

f465. The predicted product of *f465* at 83447–82050 has a similarity score of 39.2% over 278 aa to cytochrome c peroxidase from *Pseudomonas aeruginosa*, which unlike other peroxidases, has two c-type heme groups in a single polypeptide chain (44). The *f465* gene product contains three c-type heme-binding sites. In the alignment of these two proteins, one of the sites in *f465* matches the peroxidative low-potential heme of the cytochrome c peroxidase, and another matches the cytochrome-like high potential heme. The distal histidine for the peroxidative heme is also found in the *f465* gene product, at residue 415, and is preceded by 11 exactly conserved residues, LRNVALTAPYF. The distal methionine for the cytochrome-like heme is found at residue 254.

o146a. This ORF (at 1205–1645) encodes a signal-sequence-containing protein with an unusual composition, with approximately one-third of its residues either glutamine or asparagine. Several proteins in the database are listed in Table 2 as similar to the *o146a* gene product; however, these proteins are also very rich in glutamine and asparagine residues. The alignment of glutamine and asparagine residues between *o146a* and these proteins may reflect the amino acid composition rather than evolutionary or structural relationships.

Putative carbohydrate utilization genes. A cluster of nine genes is found at 156985–165034 encoding products similar to enzymes utilized in carbohydrate metabolism, including transport, dehydrogenase, kinase, mutase, and epimerase activities. A divergently transcribed gene, *f282*, encoding a putative transcriptional repressor protein is found immediately upstream. The *o231* gene product is very similar to AraD, or L-ribulose-5-phosphate-4-epimerase, which catalyzes an early step in L-arabinose utilization, with scores of 76.2% over 231 aa to the *E.coli* enzyme and 74.4% over 203 aa to that of *S.typhimurium*. The *o328* product has a score of 27.5% over 231 aa to DctP in *Rhodobacter capsulatus*, the periplasmic binding protein component of a C-4 dicarboxylate transport system (21). The *o328* product, like DctP, contains a possible signal sequence. The adjacent *o424* gene encodes a protein that could also be a transport component; it has a similarity score of 32.8% over 401 aa to an integral membrane protein from *Rhodobacter capsulatus*. The products of *o498*, *o332*, and *o220* are similar to sugar kinase, dehydrogenase, and phosphomutase proteins respectively.

Internal similarity of ORF products. Two pairs of ORFs were found in this segment with very high similarity at both the DNA and protein level. The first pair, *f378b* (*yibH*) and *f120* (*yibI*), is located between *rhsA* and *mlaA* at 185996–184495. The second pair, *f378a* and *f107*, is found about 17 kb upstream. The *f378b* (*yibH*) and *f378a* products have a similarity score of 77.2% over their entire length. The *f107* and *f120* products have a score of 75.2% over 112 aa, but 93.8% over 80 residues of the N terminus.

Signal peptides

Prokaryotic signal sequences typically contain a hydrophobic core of from 7 to 20 residues, are preceded by positively charged residues and are followed by a cleavage site for leader peptidase. These regions are referred to as the H, N, and C regions respectively. A specially-written program was used to detect signal sequences based on the weight-matrix method of von Heijne (45). Eight genes were found in this segment that had been previously identified as encoding periplasmic proteins; these genes are *ggt*, *ugpB*, *livJ*, *livK*, *nikA*, *dppA*, *xylF*, and *malS* (13, 46–51). The *ggt* gene encoding gamma-glutamyltranspeptidase extends beyond the left end of the sequence and is reported here as an N-terminal fragment. This portion of Ggt is part of the larger of two subunits cleaved from a single polypeptide after export by means of a signal sequence.

The computer search for signal peptide-like sequences found twenty-three candidates in addition to the ones already identified. Of these, at least five most likely encode membrane-bound proteins: *o146a* at 1205–1645, *livM*, *nikB*, *dppC*, and *lctP* (49, 52–55). Two others, *f110* and *f112* between 70999–70225, were previously identified as cotranscribed genes encoding soluble (cytoplasmic/periplasmic) proteins whose expression is affected by the nucleoid protein H-NS (56). The N-terminal amino acid sequences were determined for these proteins (56) and match the translation product of *f110* and *f112* beginning, respectively, 21 and 33 residues after the start codon, which is consistent with processing of a signal peptide on these proteins. Upstream of *f110* and *f112* is *f215*, another protein containing a predicted signal sequence. Also containing signal peptides are the products of *f1165*, *f368*, and *f779* between 106859–99956, which are similar to cellulose metabolizing enzymes as discussed above, and the product of *o328* at 160347–161333, which is similar to the periplasmic C-4 dicarboxylic acid binding protein encoded by *dctP* in *Rhodobacter capsulatus*.

Other genes encoding proteins with predicted signal sequences are *o208* (20013–20639), *o203* (24163–24774), *o376* (26046–27176), *f355* (44862–43795), *f498* (96196–94700), *o423* (114816–116087), *f234* (127192–126488), *f274b* (151428–150604), *o220* (162830–163492), *f246* (166121–165381), and *o427* (201064–202347). The *o498* gene product in this group may be a protease (see Table 2).

Other features

Cold-regulated and heat shock genes. Two genes in this region are known to be cold regulated, *ggt* at the left end of this segment and *cspA* at 134300–134512 (46, 57). In addition, the phenotype of some *rfaG* mutants includes cold-sensitive inhibition of cell division and sensitivity to detergents (29). Two genes are thought to be regulated by heat shock in this segment, *prlC* at 59441–57399 and *rfaD* at 208251–209183 (58, 59). Four other sigma 32 (heat shock) regulated promoters were identified as having a good match to the consensus, but no experimental evidence is currently available to confirm their activity *in vivo*. These promoters are found preceding *f419* (26037–24778), *xylA* (145016–143694), *f274b* (151428–150604), and *o157a* (158580–159053).

Transcription signals. Promoter and terminator sites were predicted using computer searches and human assessment, taking into account transcription data already in the literature wherever possible. Their arrangement indicated 90 transcription units (Figure 1). As well as the sigma 32 promoters, nine other

Figure 1. Map of the finished sequence and its features, proceeding left to right in 13 tiers. The top line shows restriction sites for ten enzymes: *AvrII* (A), *BamHI* (B), *BglII* (G), *EcoRI* (R), *EcoRV* (V), *HindIII* (H), *KpnI* (K), *NotI* (N), *PstI* (S) and *PvuII* (P). In addition, the location and orientation of Chi sites are indicated by X> and <X. Gene names are indicated for identified RNA-coding regions and open reading frames. Unidentified open reading frames are designated o# or f# for the two transcriptional orientations, where the numbers indicate the predicted peptide length. Promoters marked F, N or H are regulated by sigma factors 28, 54, and 32, respectively; unlabeled promoters are regulated by sigma 70. Other sequence determinations spanned by this segment are indicated, with their database accession numbers. The annotated sequence has been submitted to GenBank as ECOUW76 (Accession U00039).

'alternative' promoters were assigned, six by sigma 54, and three by sigma 28. Sigma 54 promoters (N) are found preceding *rpoH* (15045–14191), *o383* (49001–50152), *pit* (51901–53400), *f465* (83447–82050), *bisC* (130532–128313), and *o514* (199510–201054). Sigma 28 promoters (F) are found preceding *o127* (12246–12629), *o378* (38640–39776), and *f216* (93443–92793).

Non-coding features (all shown in Figure 1). Nineteen REP (Repeated Extragenic Palindromic) elements (60) were identified by computer search, some human assessment, and comparison with previously reported data (61, 62). No ERIC elements (enterobacterial repetitive intergenic consensus), also called IRU (intergenic repeat unit), were found. The REP element at 155279–155312 downstream of *o417*, is the same structure as the 'RIP element upstream of *avtA*' described by Oppenheim *et al.* (63) in that the spacer between the two parts of the REP element is similar to the IHF binding site consensus. Data from a gel retardation experiment confirms this relationship (63), which was also noticed by Boccard and Prentki (64). In our sequence this structure is well positioned to act as a terminator for *o417* transcription. Differences in this area are discussed below.

Two insertion sequences, IS5 and IS150, were found in this segment (65). IS5 was originally isolated as an insertion (KH100) of about 1200 bp (2.4% λ) in the *cI* gene of bacteriophage λ (66).

The computer predicted 95 static bends in the DNA greater than 72 degrees, calculated as described previously. Finally, the segment 183502–184495 (994 bp), just beyond the second repeat in the *RhsA* element, is devoid of any detectable features; we therefore designate this area as a grey hole (1).

Gene orientation and Chi sites. In previously analyzed segments it was found that two thirds of genes were oriented in the direction

of DNA replication (2–4). In this segment the rule no longer applies. Genes are arranged with equal numbers in each direction, whether measured by gene number or by total length of DNA expressed in either orientation. The close correlation between Chi site orientation and direction of translation or replication observed previously still holds true in this segment (2). Of the 50 Chi sites occurring within coding sequences, only 2 are opposed to replication and translation. Thirty Chi sites are oriented with both translation and replication and 11 are oriented with either translation or replication, but not both.

Codon usage and horizontal gene transfer

Some ORFs in this segment were seen to exhibit very poor codon usage scores when analyzed using compiled matrices of *E. coli* codon usage frequency for highly expressed genes. These ORFs generally showed higher scores when compared with the 'Phage' matrix compiled by Borodovsky, derived from the genes in class III as described by Médigue *et al.* (67). The genes in this class encode cell surface components involved in mating and motility, and host integration factors, and include genes from phage lambda, plasmids and insertion sequences. They were suggested to represent examples of horizontal gene transfer, which if true might provide a new category of analysis. Twenty genes in this segment show good scores with the 'Phage' matrix but poor scores with the *E. coli* matrix. The orfs that are associated with the *Rhs* elements in this segment are in this category, *o280b* and *o153* between 180589–182890 downstream of *rhsA*, and *o122* and *o378* between 37676–39776 downstream of *rhsB*. Another group of genes with class III-type codon usage scores are in the *rfa* cluster at the right end of this segment, which encode proteins functioning in lipopolysaccharide biosynthesis. *rfaLKZYJIBSPGQ* and the upstream gene *f290* show the class III statistics. *rfaD*, *rfaF*, and *rfaC* are the only genes in this cluster that have typical

E. coli codon usage statistics. RfaL has several hydrophobic domains (potential transmembrane domains) typical of an integral membrane protein. Klena *et al.* speculate that the role of RfaL in surface O antigen expression is somehow strain- or species-specific (68).

Of the ORFs in the insertion sequences in this segment, *f338* (IS5) at 67457–66441 shows similar scores with both matrices, whereas in IS150, *o173* and *o238* (between 134931–136300) both show class III codon usage. Other genes in this segment with class III-type codon usage are *f540* at 46850–45228, *o260* at 65550–66332, *o175* at 72625–73152, and *f274a* at 80069–79245.

Differences from published sequences

avt. The *avtA* gene encoding alanine-valine transaminase (EC-AVT, Y00490) was putatively identified by Berg and coworkers who sequenced a region downstream of *malS* and identified an open reading frame in the reverse orientation (69). The gene was characterized using transposon insertions in this region and measuring AvtA expression (69, 70). Enigmatically, two of the insertions reduced expression even though they were not in the coding region.

The ECAVT sequence differs from our determination in more than 60 places. Instead of the gene described by Berg *et al.*, we find a larger ORF in the forward orientation, *o417* at 153956–155209, that contains all the transposon sites that affected AvtA expression. Analysis of codon usage in ECAVT suggested that there may be frameshift errors in the previous determination of the sequence.

The results of computer searches suggest the possibility of transaminase activity for the *o417* product. A low similarity score, 25.6% over 115 aa, is seen for a protein with transamidating activity, CobC from *Pseudomonas denitrificans* (COBC_PSEDE) (71), and very low similarity, close to background levels, to several aminotransferases. Both CobC and the *o417* gene product contain a close match to the PROSITE motif PS00105 AA__TRANSFER__CLASS__1 (aminotransferases class I pyridoxal-phosphate attachment site) with 9 out of 10 conserved or partially conserved sites in a 14 residue pattern. The putative ECAVT ORF did not show these features; however, we feel that a conclusive assignment of *avtA* would be premature at this time and requires more experimental data.

dppF. The published sequence (L08399) has 283 bases missing relative to our determination. On consultation with the author it was noted that the missing sequence is bounded by *HincII* sites and is thought to have been lost during subcloning with this enzyme prior to sequencing.

E. coli genes. In this segment of *E. coli* we found two nearly perfect matches to sequences in GenBank described as putative human genes. Human chromosome 8 STS UT5352 (GenBank L18630) is 95.7% identical over its whole length to bases 153564–153995. Human cDNA clone FB20A1 (GenBank T03011) is 93.5% identical over its whole length with bases 150047–150469. Given the evolutionary distance between *E. coli* and human, we suggest that *E. coli* DNAs were accidentally sequenced as human DNA.

Updates

There are no new changes to report in our previously determined sequences.

ACKNOWLEDGEMENTS

This is paper 3400 from the Laboratory of Genetics, supported by award HG000301 from the NIH Human Genome Project. We thank M. Kröger, K. Rudd, A. Link and G. Church for their continuing help with databases; A. Oppenheim, M. Manson, C. Berg, and K. Rudd were consulted on points of difference. For software development and assistance we thank M. Borodovsky, G. Bouriakov, D. Argentar, J. Shavlik, and the programming staff of DNASTAR. Excellent technical support was provided by B. Fritz, K. Kadner, M. Maguire, R. Mikkelsen, D. Rose, A. Schmidt, E. Sommers, S. Subramanian, R. Talley. Thanks also to N. Peterson for administration. Finally we thank our University of Wisconsin–Madison undergraduates and recent graduates for their excellent teamwork: H. Abernathy, Z. Abu-Hassan, S. Ahmad, A. Azman, A. Broah, J. Cairns, J. Champ, H. Cheng, A. Copeland, W. Davis, K. Day, H. Demrow, D. Duescher, J. Ehley, E. Grotbeck, L. Hammes, J. Hobbes, D. Hornung, M. Kamari, J. Kamaruddin, J. Katcha, H. Kijenski, Heather Kirkpatrick, Heidi Kirkpatrick, R. Mohamad, A. Mohamed, E. Mohd-Nazir, Z. Othman, M. Polka, S. Rhode, T. Ryan, P. Scherkenbach, J. Schmidt, L. Schroeder, K. Streif, W. Wilson, C. Whipperman, and R. Williams.

REFERENCES

- Daniels, D.L., Plunkett, G. III, Burland, V. and Blattner, F.R. (1992) *Science* **257**, 771–778.
- Burland, V., Plunkett, G. III, Daniels, D.L. and Blattner, F.R. (1993) *Genomics* **16**, 551–561.
- Plunkett, G. III, Burland, V., Daniels, D.L. and Blattner, F.R. (1993) *Nucleic Acids Res.* **21**, 3391–3398.
- Blattner, F.R., Burland, V., Plunkett, G. III, Sofia, H.J. and Daniels, D.L. (1993) *Nucleic Acids Res.* **21**, 5408–5417.
- Borodovsky, M. and McIninch, J. (1993) *Computers & Chemistry* **17**, 123–133.
- Bachmann, B.J. (1990) *Microbiol. Rev.* **54**, 130–197.
- Silver, S., Ji, G., Bröer, S., Dey, S., Dou, D. and Rosen, B.P. (1993) *Mol. Microbiol.* **8**, 637–642.
- Ji, G. and Silver, S. (1992) *J. Bacteriol.* **174**, 3684–3694.
- Greer, S. and Perham, R.N. (1986) *Biochemistry* **25**, 2736–2742.
- Oden, K.L., Gladysheva, T.B. and Rosen, B.P. (1994) *Mol. Microbiol.* **12**, 301–306.
- Rosen, B.P., Dey, S., Dou, D., Kaur, P. and Ksenzenko, M. (1994) *Antonie van Leeuwenhoek*. In press.
- Boos, W., Ehmann, U., Forkl, H., Klein, W., Rimmel, M. and Postma, P. (1990) *J. Bacteriol.* **172**, 3450–3461.
- Sumiya, M. and Henderson, P.J.F. (1989) *Biochemical Society Transactions* **17**, 553–554.
- Batt, C.A., Bodis, M.S., Picataggio, S.K., Claps, M.C., Jamas, S. and Sinskey, A.J. (1985) *Can. J. Microbiol.* **31**, 930–933.
- Rosenfeld, S.A., Stevis, P.E. and Ho, N.W.Y. (1984) *Mol. Gen. Genet.* **194**, 410–415.
- David, J.D. and Wiesmeyer, H. (1970) *Biochim. Biophys. Acta* **201**, 497–499.
- Sprague, G.F., Jr., Bell, R.M. and Cronan, J.E., Jr. (1975) *Mol. Gen. Genet.* **143**, 71–77.
- Elvin, C.M., Hardy, C.M. and Rosenberg, H. (1987) In Torriani-Gorini, A., Rothman, F.G., Silver, S., Wright, A. and Yagil, E. (eds.), *Phosphate Metabolism and Cellular Regulation in Microorganisms*. ASM Publications, Washington, D.C., pp. 156–158.
- Kay, W.W., Sweet, G.D., Widenhorn, K. and Somers, J.M. (1987) In Rosen, B.P. and Silver, S. (eds.), *Ion Transport in Prokaryotes*. Academic Press, San Diego, CA, pp. 269–302.
- Lo, T.C.Y. and Bewick, M.A. (1978) *J. Biol. Chem.* **253**, 7826–7831.
- Shaw, J.G., Hamblin, M.J. and Kelly, D.J. (1991) *Mol. Microbiol.* **5**, 3055–3062.
- Rodolakis, A., Casse, F. and Starka, J. (1974) *Mol. Gen. Genet.* **130**, 177–181.
- Klein, J.R., Henrich, B. and Plapp, R. (1991) *Mol. Gen. Genet.* **230**, 230–240.
- Klein, J.R. and Plapp, R. (1992) *J. Bacteriol.* **174**, 3828–3829.

25. Ma, D., Cook, D.N., Alberti, M., Pon, N.G., Nikaido, H. and Hearst, J.E. (1993) *J. Bacteriol.* **175**, 6299–6313.
26. Xu, J. and Bertrand, K.P. (unpublished). GenBank accession M94248.
27. Xu, J. and Bertrand, K.P. (unpublished). GenBank accession M96848.
28. Pouyssegur, J. and Stoeber, F. (1974) *J. Bacteriol.* **117**, 641–651.
29. Schnaitman, C.A., Parker, C.T., Klena, J.D., Pradel, E.L., Pearson, N.B., Sanderson, K.E. and MacLachlan, P.R. (1991) *J. Bacteriol.* **173**, 7410–7411.
30. Wong, H.C., Fear, A.L., Calhoon, R.D., Eichinger, G.H., Mayer, R., Amikam, D., Benziman, M., Gelfand, D.H., Meade, J.H., Emerick, A.W., Bruner, R., Ben-Bassat, A. and Tal, R. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 8130–8134.
31. Saxena, I.M., Lin, F.C. and Brown, R.M., Jr. (1990) *Plant Mol. Biol.* **15**, 673–683.
32. Saxena, I.M., Lin, F.C. and Brown, R.M., Jr. (1991) *Plant Mol. Biol.* **16**, 947–954.
33. Atkinson, E.M. and Long, S.R. (1992) *Mol. Plant–Microbe Interact.* **5**, 439–442.
34. Henrissat, B., Claeysens, M., Tomme, P., Lemesle, L. and Mornon, J.-P. (1989) *Gene* **81**, 83–95.
35. Mayer, R., Ross, P., Weinhouse, H., Amikam, D., Volman, G., Ohana, P., Calhoon, R.D., Wong, H.C., Emerick, A.W. and Benziman, M. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 5472–5476.
36. Priefert, H., Krüger, N., Jendrossek, D., Schmidt, B. and Steinbüchel, A. (1992) *J. Bacteriol.* **174**, 899–907.
37. Mallinder, P.R., Pritchard, A. and Moir, A. (1992) *Gene* **110**, 9–16.
38. Murgola, E.J. and Adelberg, E.A. (1970) *J. Bacteriol.* **103**, 178–183.
39. Lapointe, J. and Delcuve, G. (1975) *J. Bacteriol.* **122**, 352–358.
40. Lapointe, J. and Söll, D. (1972) *J. Biol. Chem.* **247**, 4982–4985.
41. Jahn, D. (1992) *FEBS Lett.* **314**, 77–80.
42. Gendron, N., Breton, R., Champagne, N. and Lapointe, J. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 5389–5392.
43. Osorio, A.V., Camarena, L., Salazar, G., Noll-Louzada, M. and Bastarrachea, F. (1993) *Mol. Gen. Genet.* **239**, 400–408.
44. Rönnberg, M., Kalkkinen, N. and Ellfolk, N. (1989) *FEBS Lett.* **250**, 175–178.
45. von Heijne, G. (1986) *Nucleic Acids Res.* **14**, 4683–4690.
46. Suzuki, H., Kumagai, H., Echigo, T. and Tochikura, T. (1989) *J. Bacteriol.* **171**, 5169–5172.
47. Overduin, P., Boos, W. and Tommassen, J. (1988) *Mol. Microbiol.* **2**, 767–775.
48. Rahmanian, M., Claus, D.R. and Oxender, D.L. (1973) *J. Bacteriol.* **116**, 1258–1266.
49. Navarro, C., Wu, L.-F. and Mandrand-Berthelot, M.-A. (1993) *Mol. Microbiol.* **9**, 1181–1191.
50. Manson, M.D., Blank, V., Brade, G. and Higgins, C.F. (1986) *Nature* **321**, 253–256.
51. Freundlieb, S. and Boos, W. (1986) *J. Biol. Chem.* **261**, 2946–2953.
52. Tommassen, J., Eiglmeier, K., Cole, S.T., Overduin, P., Larson, T.J. and Boos, W. (1991) *Mol. Gen. Genet.* **226**, 321–327.
53. Adams, M.D., Wagner, L.M., Graddis, T.J., Landick, R., Antonucci, T.K., Gibson, A.L. and Oxender, D.L. (1990) *J. Biol. Chem.* **265**, 11436–11443.
54. Abouhamad, W.N., Manson, M., Gibson, M.M. and Higgins, C.F. (1991) *Mol. Microbiol.* **5**, 1035–1047.
55. Dong, J.M., Taylor, J.S., Latour, D.J., Iuchi, S. and Lin, E.C.C. (1993) *J. Bacteriol.* **175**, 6671–6678.
56. Yoshida, T., Ueguchi, C., Yamada, H. and Mizuno, T. (1993) *Mol. Gen. Genet.* **237**, 113–122.
57. Goldstein, J., Pollitt, N.S. and Inouye, M. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 283–287.
58. Conlin, C.A. and Miller, C.G. (1993) *J. Bacteriol.* **175**, 5731–5732.
59. Raina, S. and Georgopoulos, C. (1991) *Nucleic Acids Res.* **19**, 3811–3819.
60. Yang, Y. and Ames, G.F.-L. (1990) In Drlica, K. and Riley, M. (eds.), *The Bacterial Chromosome*. American Society for Microbiology, Washington, D.C., pp. 211–225.
61. Gøberdhan, P.D., Rudd, K.E., Morgan, M. K., Bayat, H. and Ames, G.F.-L. (1992) *J. Bacteriol.* **174**, 4583–4593.
62. Gilson, E., Saurin, W., Perrin, D., Bachellier, S. and Hofnung, M. (1991) *Nucleic Acids Res.* **19**, 1375–1383.
63. Oppenheim, A.B., Rudd, K.E., Mendelson, I. and Teff, D. (1993) *Mol. Microbiol.* **10**, 113–122.
64. Boccard, F. and Prentki, P. (1993) *EMBO J.* **12**, 5019–5027.
65. Birkenbihl, R.P. and Vielmetter, W. (1989) *Mol. Gen. Genet.* **220**, 147–153.
66. Blattner, F.R., Flandt, M., Hass, K.K., Twose, P.A. and Szybalski, W. (1974) *Virology* **62**, 458–471.
67. Médigue, C., Rouxel, T., Vigier, P., Hénaut, A. and Danchin, A. (1991) *J. Mol. Biol.* **222**, 851–856.
68. Klena, J.D., Pradel, E. and Schnaitman, C.A. (1992) *J. Bacteriol.* **174**, 4746–4752.
69. Liu, L., Whalen, D.A. and Berg, C.M. (1987) *Nucleic Acids Res.* **22**, 9461–9469.
70. Wang, M.-D., Liu, L., Wang, B. and Berg, C.M. (1987) *J. Bacteriol.* **169**, 4228–4234.
71. Crouzet, J., Cauchois, L., Blanche, F., Debussche, L., Thibaut, D., Rouyez, M.-C., Rigault, S., Mayaux, J.-F. and Cameron, B. (1990) *J. Bacteriol.* **172**, 5968–5979.
72. Berlyn, M.B. and Letovsky, S. (1992) *Nucleic Acids Res.* **20**, 6143–6151.
73. Fuchs, R. (1991) *Comput. Applic. Biosci.* **7**, 105–106.