Analysis of the *Escherichia coli* genome VI: DNA sequence of the region from 92.8 through 100 minutes

Valerie Burland*, Guy Plunkett III, Heidi J. Sofia⁺, Donna L. Daniels and Frederick R. Blattner

Laboratory of Genetics, University of Wisconsin, 445 Henry Mall, Madison, WI 53706, USA

Received March 31, 1995; Revised and Accepted May 12, 1995

GenBank accession no. U14003

ABSTRACT

The 338.5 kb of the *Escherichia coli* genome described here together with previously described segments bring the total of contiguous finished sequence of this genome to >1 Mb. Of 319 open reading frames (ORFs) found in this 338.5 kb segment, 147 (46%) are potential new genes. The positions of several genes which had been previously located here by mapping or partial sequencing have been confirmed. Several ORFs have functions suggested by similarities to other characterised genes but cannot be assigned with certainty. Fifteen of the ORFs of unknown function had been previously sequenced. Eight transfer RNAs are encoded in the region and there are two grey holes in which no features were found. The attachment site for phage P4 and three insertion sequences were located. The region was also analysed for chi sites, bend sites, REP elements and other repeats. A computer search identified potential promoters and tentative transcription units were assigned. The occurrence of the rare tetramer CTAG was analysed in 1.6 Mb of contiguous *E. coli* sequence. Hypotheses addressing the rarity and distribution of CTAG are discussed.

INTRODUCTION

The 338 524 bp described here is the sixth segment of the *Escherichia coli* genome sequenced by this project, and completes 1.06 Mb of contiguous *E. coli* sequence from a single strain (1-5). In the course of this project we have seen many strain differences resulting in different coding potential; two are described here (*sugES* and *phnE*). These emphasize the importance of sequencing a single genome, and support the idea that a sequence mosaic derived from many strains will not make genetic sense and cannot be definitive.

The structure of this region of the genome, and the arrangement of features contained conforms to that observed in previously analysed regions; that of very densely packed genes with only two grey holes (1). Only one new gene was identified by protein similarity searches, and the number of previously sequenced but

still unidentified genes found by database searching is increasing (15 in this segment).

Our sequence overlaps (at *o228* adjacent to *thrA*) the segment from 0 to 4.1 minutes which was determined elsewhere (6,7), raising the question of the precise physical definition of the 0 minute point.

MATERIALS AND METHODS

The sequence was all newly determined from 30 lambda clones of the MG1655 encyclopedia (8) by a mixed strategy of random then directed data collection, details described previously (1,9). The sequence was carefully compared with previously published determinations and differences resolved wherever possible. Computer programs were used to find putative coding sequences, promoters, terminators, repeat sequences, chi sites, bend sites and signal peptides. Database searches placed previous DNA determinations in context within our sequence and suggested functions for new ORFs by matches to known proteins. The similarity index resulting from the amino acid (aa) sequence alignment indicates % identity corrected for gaps. Software and other analytical processes have been described previously (1-5).

In the course of this work we have investigated methods for systematic resolution of compressions observed when anomalous mobility of DNA fragments with GC rich 3' ends prevents correct reading of the sequence. This is a significant problem because *E. coli* is GC-rich (51%) relative to many of the other genomes being sequenced, because gene-finding is affected, and most (85%) of this genome is coding sequence. Although use of dITP substituted for dGTP in the radiolabelled Sequenase reactions resolves most compressions, pauses appear, where the polymerase fails to proceed. A more satisfactory technique is to use fluorescent labelled dideoxy terminators to label the 3' ends of the DNA fragments in *Taq* DNA polymerase thermal cycling reactions, and collect the data with an ABI 373A sequencer. dITP is also used in these reactions, but paused fragments fail to label and are not detected. In collaboration with Applied Biosystems, a set of 12 different compressions chosen to test the method were all successfully resolved (manuscript in preparation). Fluorescent labelled primers have no effect on compressions.

* To whom correspondence should be addressed at present address: FMC Bioproducts, 191 Thomaston Street, Rockland, ME 04856, USA

⁺Present address: National Center for Biotechnology Information, National Library of Medicine/NIH, Building 38A, Room 8N805, 8600 Rockville Pike, Bethesda, MD 20894, USA

By the time the last part of this sequence was being collected, a systematic method was in place in which random data was collected using radiolabelled Sequenase reactions with dGTP and the sequences were assembled in the computer. Then a minimal spanning set of clones was selected by a specially written program, and these were sequenced with 3' fluorescent label as described above. Adding this data to the assembled sequence has significantly reduced editing time and enabled resolution of nearly all the compressions.

RESULTS AND DISCUSSION

A map of the 338 533 bp of sequence and the features found within it is shown in Figure 1. The finished sequence, deposited in GenBank as ECOUW93 with the accession number U14003 was first analysed for open reading frames, denoted *o* or *f* (ORF or FRO) for the two orientations. Table 1 lists the genes and ORFs with their coordinates and their predicted physical characteristics. Previously sequenced or partially sequenced genes were identified by searching the sequence against DNA databases and differences were confirmed by checking the primary data. Searches of the protein databases gave clues to the functions of many ORFs and led to the identification of several genes genetically mapped to this area. Results of the protein database searches are shown in Table 2. Database entries are referred to by their locus names or accession numbers.

Mapped genes identified

***mgtA*.** An ATP-dependent magnesium transport locus had been mapped to this genomic segment (10), and the analogous gene in *Salmonella typhimurium*, *mgtA* has been sequenced (STU07483). The MgtA amino acid (aa) sequence is 88.4% similar over its whole length to an ORF of 898 aa at 158 454–161 150. The new gene is located ~125 kb (~3 minutes) from the original mapped site. In addition the DNA sequences of the two genes are 73% similar over 2696 bases. These convincing similarities indicate the designation of *o898* as *mgtA*.

***gntP*.** A secondary gluconate transport system for *E.coli* was originally described as *gntS*, cotransducible with *fdp* (11,12). More recently, genetic evidence and partial sequence data has identified a gluconate kinase (*gntV*) between the *fec* operon and *valS* near 97 minutes on the 1990 genetic map (K. Rudd, personal communication; T. Conway, personal communication). Conway also reports genetic evidence of a permease activity mapping between *fecI* and *uxuA* (personal communication), ~1 minute clockwise from *gntV*.

Three ORF products in this area show similarity to the *Bacillus subtilis* gluconate permease *GntP*. The candidate ORFs are *f439* at 182 041–183 360, *f388* at 211 686–212 852 and *f477* at 240 785–242 128. The protein similarities are 49.1, 57.1 and 36.5%, respectively. These ORF products are also similar to each other at both the protein and DNA levels. Candidate *f447* is the closest in size to the *B.subtilis* gene though its aa sequence match is the poorest of the three. Its location, between the *fim* and *uxu* operons corresponds with the permease studied by Conway.

Candidate *f439* is only two ORFs from *gntV*. The two interposed ORFs are *f254* and *f343a*, whose respective products are similar to 2-keto-3-gluconate reductase from *Erwinia*, (41.8%), and sorbitol dehydrogenases from several organisms (~29%). The suggestion of a possible operon is mildly supported by *f332*, adjacent to the proposed *gntP* which shows 25.1% similarity over 329 aa with *lacI*

family regulator proteins. In *B.subtilis*, the gluconate operon comprises a repressor, a kinase, and a permease gene in that order, followed by a fourth gene of unknown function (13).

Candidate *f388* is shorter than the *B.subtilis* gene (448 aa) and the match, though good, extends over only 355 aa. The flanking ORFs are *f266* on the 5' side, whose product is 26% similar to *sorC*, a regulatory protein from *Klebsiella pneumoniae*. On the 3' side is *f655* coding a product 26.8% similar to *ilvD* of *E.coli* and 28.7% similar to *edd* of *Zymomonas mobilis*, a phosphogluconate dehydratase.

None of these candidates correspond with the original map location of *gntS*, and none of the unidentified ORFs near *fdp* show any similarity to the *B.subtilis* *gntP*. However, the products of a group of four ORFs adjacent to *fdp*, all show similarity (29–52%) with high affinity transport proteins for several other sugars. Although there is no evidence of a role in gluconate metabolism beyond the early mapping data, these matches are suggestive of a new operon.

***nadR*.** The repressor of the NAD synthesis genes has been well characterised in *S.typhimurium* (14,15), and its gene, *nadR*, has been sequenced (16). *o417* at 318 125–319 378 is 95% similar to the *S.typhimurium* gene. A partial sequence of an *E.coli* ORF (SwissProt P27278) also matches both the *E.coli* and salmonella genes. This very strong similarity enables us to designate *o417* as *nadR*. In *S.typhimurium*, this gene also has a role in nucleotide transport. However a literature search revealed no new studies on this gene of *E.coli* in the past decade.

***vacB*.** The chromosomally located *vacB* gene of *Shigella flexnerii* regulates virulence by interaction between its product and gene products encoded by the large plasmid which are required for enteroinvasive functions (17). The *vacB* gene was also detected in enteroinvasive isolates of *E.coli* (17). The end of a partially sequenced *E.coli* K-12 ORF downstream of *purA* (18) matched the *Shigella* sequence and we now confirm that the full gene in K-12, *o827* at 97 440–99 923, is highly similar at the DNA (93%) and protein level (89%). The role of *vacB* in non-pathogenic strains is not known.

***uxu* operon.** The enzymes encoded in the *uxu* region of *E.coli* are well studied and are involved in the pathway of glucuronate degradation (19,20). The regulatory region was sequenced (21) and we found these 434 bases to be immediately upstream of the ORFs *o394*, *o471* and *o257* at 242 468–246 180, overlapping the start of *o394*. We therefore identify these ORFs as *uxuA*, *B* and *R*, respectively. The identification is supported by identity of the restriction maps for the whole operon (22) and mapping data (23), but the whole operon is in the opposite orientation to that shown on the current genetic map (10).

***cst*.** Of the many that are induced by carbon starvation in *E.coli*, three cAMP-dependent genes have been identified: *cstA* maps at 14 minutes and has been sequenced, and *cstB* (not sequenced) maps at 93 minutes (24,25). The third locus is unmapped. Two new ORFs located near 99 minutes, rather distant from the mapped site of *cstB*, could be this third locus. *f721* and the adjacent small ORF *f67* at 279 709–282 127 in this sequence encode products that are similar to the products of three adjacent ORFs, *cstA*, *ybdC* and *ybdD*. *F721* and *CstA* (562 aa) are 66% similar over 554 aa starting at the N-terminus of *F721*; the predicted product of the putative ORF *ybdC* is 42% similar to the C-terminal 168 aa of *F721*. The open reading frames of *cstA* and

Table 1. Genes and predicted characteristics of deduced protein products

[illegible]

Table 1. continued

<i>f442 (y41)</i>	194265 < 192937	ATG..TAA	50.4	442	9.7
<i>o425</i>	194832 > 196109	TTG..TGA	46.6	425	9.4
<i>o377</i>	196106 > 197239	ATG..TAA	41.9	377	6.4
<i>f124</i>	197834 < 197460	ATG..TAA	13.7	124	7.0
<i>o67</i>	197740 > 197943	ATG..TAA	7.1	67	6.2
<i>o100</i>	197995 > 198297	GTG..TAG	11.6	100	10.1
<i>is30</i>	199451 < 198300	ATG..TGA	44.3	383	10.9
<i>o198</i>	199792 > 200388	ATG..TGA	22.3	198	10.2
<i>o137</i>	200554 > 200967	ATG..TAA	15.0	137	8.3
<i>fecE</i>	202291 < 201524	ATG..TAG	28.2	255	8.2
<i>fecC</i>	203248 < 202292	ATG..TAA	34.1	318	10.0
<i>fecB</i>	204243 < 203245	ATG..TGA	34.9	332	9.5
<i>fecA</i>	205148 < 204240	GTG..TGA	33.4	302	8.9
<i>fecR</i>	207511 < 205187	ATG..TGA	85.3	774	5.7
<i>fecI</i>	208551 < 207598	ATG..TAA	35.5	317	9.0
<i>insA</i>	209069 < 208548	ATG..TGA	19.5	173	6.0
<i>f266</i>	209361 > 209636	GTG..TAA	9.9	91	10.7
<i>f388</i>	210972 < 210172	GTG..TAA	29.1	266	7.8
<i>f655</i>	212852 < 211686	ATG..TGA	41.1	388	8.7
<i>f319</i>	214926 < 212959	ATG..TGA	70.1	655	6.0
<i>f262a</i>	215896 < 214937	ATG..TGA	34.9	319	5.0
<i>f260</i>	216635 < 215847	ATG..TGA	29.9	262	8.7
<i>f210</i>	217720 < 216938	ATG..TGA	29.3	260	6.5
<i>f143</i>	218369 < 217737	ATG..TAA	23.2	210	6.1
<i>f268a</i>	218812 < 218381	ATG..TAA	15.6	143	5.1
<i>f437</i>	219749 < 218943	ATG..TAA	29.4	268	5.6
<i>f383</i>	221075 < 219762	ATG..TAA	46.7	337	8.9
<i>f248</i>	222513 < 221362	GTG..TGA	41.7	383	5.3
<i>f181</i>	224015 < 223269	ATG..TAA	27.4	248	4.8
<i>o338</i>	224616 < 224071	ATG..TGA	20.0	181	6.2
<i>f326b</i>	225847 < 226863	ATG..TAA	38.0	338	7.0
<i>f404</i>	228426 < 227446	ATG..TGA	36.9	326	6.3
<i>f241</i>	229705 < 228491	ATG..TAA	43.7	404	8.2
<i>finB (pil)</i>	230342 < 229617	GTG..TAG	28.3	241	8.7
<i>finE</i>	231789 < 232391	ATG..TAG	23.0	200	10.2
<i>finA (pilA)</i>	232869 < 233465	GTG..TGA	23.1	198	9.9
<i>finI</i>	233947 < 234495	ATG..TAA	18.1	182	5.0
<i>finC</i>	234452 < 235099	GTG..TAA	23.6	215	10.1
<i>finD</i>	235136 < 235861	GTG..TAA	26.7	241	9.4
<i>finF</i>	235929 < 238565	ATG..TAA	96.5	878	6.9
<i>finG</i>	238575 < 239105	ATG..TAA	18.7	176	7.2
<i>finH</i>	239118 < 239621	ATG..TGA	17.3	167	5.5
<i>uzaA</i>	239641 < 240543	ATG..TAA	31.5	300	7.7
<i>uzaB</i>	242128 < 240785	ATG..TAA	47.1	447	7.1
<i>uzaR</i>	242468 < 243652	ATG..TAA	44.8	394	5.4
<i>f276</i>	243733 < 245193	ATG..TAA	53.6	486	5.3
<i>o133b</i>	245408 < 246181	ATG..TGA	29.3	257	4.9
<i>f303</i>	247152 < 246522	ATG..TAA	30.6	276	6.0
<i>f390a</i>	247816 < 248217	GTG..TAA	15.2	133	6.9
<i>f153</i>	249121 < 248210	ATG..TGA	34.7	303	5.8
<i>f231</i>	250358 < 249186	ATG..TAA	41.1	390	5.0
<i>o218</i>	250832 < 250371	ATG..TAA	16.2	153	7.2
<i>f392</i>	251524 < 250829	ATG..TGA	24.2	231	5.1
<i>f323</i>	251660 < 252316	GTG..TAA	24.6	218	9.6
<i>f257</i>	253507 < 252329	ATG..TGA	41.4	392	9.4
<i>f390b</i>	254546 < 253575	ATG..TAA	36.0	323	7.0
<i>f426</i>	255527 < 254754	GTG..TGA	27.4	257	6.5
<i>f410</i>	256703 < 255531	GTG..TAA	43.6	390	5.2
<i>o103</i>	258078 < 256798	ATG..TAG	48.2	426	9.1
<i>o186</i>	259351 < 258119	ATG..TGA	44.7	410	10.2
<i>f470</i>	259830 < 260141	ATG..TAG	12.1	103	5.8
<i>o54</i>	260190 < 260750	ATG..TAA	21.8	186	6.2
<i>o521</i>	262406 < 260994	ATG..TAA	53.1	470	8.2
<i>o109b</i>	262583 < 262747	ATG..TGA	6.7	54	11.2
<i>o290</i>	263198 < 264763	ATG..TGA	58.6	521	6.2
<i>mcrD</i>	264967 < 265296	GTG..TGA	12.4	109	4.8
<i>mcrC</i>	265350 < 266222	ATG..TAA	33.2	290	9.2
<i>mcrB (rglB)</i>	266426 < 267688	ATG..TGA	48.3	420	5.5
<i>f132</i>	268791 < 267745	GTG..TAA	40.6	348	9.0
<i>hsdS (has)</i>	270188 < 268791	ATG..TAG	53.9	465	5.7
<i>hsdM (has, rm)</i>	270730 < 270332	GTG..TAA	14.6	132	9.4
<i>hsdR (has, rm)</i>	272295 < 270901	ATG..TGA	51.4	464	9.6
<i>mrr</i>	273881 < 272292	ATG..TGA	59.3	529	5.0
<i>f284 (yjiA)</i>	277648 < 274082	ATG..TGA	136.1	1188	5.8
<i>f67</i>	277782 < 278696	ATG..TGA	33.5	304	5.8
<i>f721</i>	279596 < 278742	GTG..TAA	32.0	284	4.4
<i>tsr (cheD)</i>	279913 < 279710	ATG..TAA	7.7	67	8.5
<i>f192</i>	282128 < 279963	ATG..TAA	77.9	721	8.7
<i>f261b</i>	282491 < 284146	ATG..TAA	59.4	551	4.8
<i>f268b</i>	284773 < 284195	ATG..TAA	20.5	192	8.6
<i>o345</i>	285555 < 284770	GTG..TGA	28.5	261	8.2
<i>mdbB</i>	286576 < 285770	ATG..TAA	31.4	268	6.9
<i>f165 (yjiA)</i>	286807 < 287844	GTG..TGA	37.1	345	6.3
<i>dnaC (dnaD)</i>	290234 < 289982	GTG..TAA	84.1	750	6.4
<i>dnaT</i>	291024 < 290527	ATG..TGA	17.5	165	5.7
<i>f108 (yjiB)</i>	291807 < 291070	ATG..TAA	27.9	245	9.0
<i>f277</i>	292349 < 291810	ATG..TAA	19.5	179	5.1
<i>o241</i>	292782 < 292456	ATG..TAA	11.9	108	10.2
<i>o225a</i>	293753 < 292920	ATG..TGA	30.5	277	8.5
<i>f262b</i>	294309 < 295034	ATG..TAG	27.0	241	7.8
<i>leuV</i>	294992 < 295669	ATG..TAA	25.6	225	9.5
<i>leuP</i>	296495 < 295707	ATG..TGA	30.1	262	6.3
<i>leuQ</i>	296997 < 296911	(RNA)	87	bp	
<i>f343b</i>	297118 < 297032	(RNA)	87	bp	
<i>holD</i>	297233 < 297147	(RNA)	87	bp	
<i>rimI</i>	298532 < 297501	ATG..TAA	37.6	343	6.3
<i>o225 (yjiG)</i>	298635 < 299048	ATG..TGA	15.2	137	5.6
	299017 < 299502	ATG..TGA	18.2	161	4.8
	299477 > 300154	ATG..TGA	25.3	225	4.7
<i>prfC (tas)</i>	300245 > 301834	ATG..TAA	59.6	529	5.8
<i>osmY</i>	302227 > 302832	ATG..TAA	21.1	201	6.6
<i>o357</i>	303242 > 304315	GTG..TGA	39.8	357	8.5
<i>o211</i>	304456 > 305091	GTG..TAG	23.7	211	7.0
<i>f287</i>	306374 < 305511	ATG..TGA	31.5	287	6.6
<i>f516</i>	307896 < 306346	ATG..TAA	58.0	516	5.9
<i>deoC (dra)</i>	308154 > 308933	ATG..TAA	27.7	259	5.5
<i>deoA (tpp)</i>	309060 > 310382	TTG..TAA	47.2	440	5.1
<i>deoB (drn)</i>	310434 > 311657	ATG..TGA	44.4	407	5.1
<i>deoD (pup)</i>	311715 > 312434	ATG..TAA	26.0	239	5.4
<i>o443 (yjiJ)</i>	312601 > 313932	ATG..TAA	49.8	443	6.2
<i>lplA (f338; yjiF)</i>	314949 < 313933	ATG..TAG	37.9	338	5.8
<i>smg</i>	315621 < 314977	ATG..TGA	24.1	214	8.7
<i>serB</i>	315727 > 316695	ATG..TAA	35.0	322	5.5
<i>sms</i>	316744 > 318126	GTG..TAA	49.5	460	7.1
<i>nadR</i>	318126 > 319379	ATG..TAA	48.2	417	5.4
<i>f579</i>	321355 < 319616	GTG..TAA	65.0	579	5.3
<i>slr</i>	321539 > 323503	ATG..TGA	74.4	654	8.5
<i>thrR</i>	323593 > 323919	ATG..TGA	12.4	108	5.4
<i>f173</i>	324587 < 324066	ATG..TAA	18.6	173	5.9
<i>o215b</i>	324630 < 325277	ATG..TAA	24.1	215	5.5
<i>rob</i>	326143 < 325274	ATG..TAA	33.1	289	7.0
<i>creA</i>	326354 < 326827	ATG..TAA	17.1	157	9.0
<i>creB</i>	326840 < 327529	ATG..TAA	26.1	229	6.4
<i>creC (phoM)</i>	327529 > 328953	ATG..TAG	52.2	474	9.2
<i>creD (cet)</i>	329011 > 330363	ATG..TAA	49.8	450	5.5
<i>dye (arcA)</i>	331139 < 330423	ATG..TAA	27.3	238	5.2
<i>o46</i>	331235 > 331375	ATG..TAA	4.9	46	8.2
<i>o228b</i>	331775 > 332461	ATG..TAA	25.3	228	5.7
<i>thrL</i>	332675 > 332740	ATG..TGA	2.1	21	11.0
<i>thrA</i>	332822 > 335284	ATG..TGA	89.1	820	5.5
<i>thrB</i>	335286 > 336218	ATG..TAA	33.6	310	5.5
<i>thrC</i>	336219 > 337505	ATG..TAA	47.1	428	5.2
<i>f121p</i>	338533 < 338168	..TAA	—	121	—

*Gene names are those used in the most recent *E. coli* genetic map (10), unless the gene is not present on that map or if another name has since become the generally accepted one; alternate gene names are indicated in parentheses. The 'p' in *f132p* and *f121p* indicates that the ORF is partial, extending beyond the sequence presented here.

*Mapped genes have been assigned Site numbers in a database maintained by the *E. coli* Genetic Stock Center (45) (M. Berlyn, personal communication). In a few cases the site number was assigned to an alternate name for a gene; those site numbers are listed in parentheses.

ybdC overlap by 136 bases (GenBank no. X52904) and examination of the codon usage in this region suggests there may be a frameshift error in the area of overlap. If this is the case then the *cstA* gene is much closer in size to *f721* than appears. The third ORF in the group, *ybdD*, is the same length as *f67*, and their predicted products are 72% similar. The presence at both sites of the small ORF with such high similarity suggests a functional relationship. However, we note that the highly similar portions of F721 and CstA are a series of membrane-spanning domains, and the match is weaker between the C-terminal region and YbdC. The possibility remains that F721 may be a transporter of a similar type but not involved in carbon metabolism.

Suggested functions

In some cases matches to proteins in the databases may suggest function, but there is no supporting evidence to warrant a new gene designation. The following cases are examples.

The predicted product of *o228* (114 532–115 218) is similar to the AraD proteins of both *E. coli* and *S. typhimurium* (59.9% over the whole aa sequence). In this case, the DNA sequences are also 57.8% similar (*o228* and *araD* of *E. coli*) thus a gene duplication or horizontal transmission is suggested.

f154 and *nrdD*. The product of *f154* at 150 729–151 193, located next to *nrdD* (encoding anaerobic ribonucleoside-triphosphate reductase), is 46.4% similar to the product of T4 gene 55.9. In the T4 genome, gene 55.9 is located near the gene encoding the phage anaerobic ribonucleoside-triphosphate reductase (*sunY* or *nrdD*;

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>f510</i>		PS00211 (P) ABC_TRANSPORTER [ABC transporters family signature]	100% match
		PS00017 (P) ATP_GTP_A [ATP/GTP-binding site motif A (P-loop)]	100% match
<i>f311</i>	311 aa	PS00017 (P) ATP_GTP_A [ATP/GTP-binding site motif A (P-loop)]	100% match
		RBSB_ECOLI (S) 296 <i>Escherichia coli</i> d-ribose-binding periplasmic protein precursor	32.3% 304 aa
		RBSB_SALTY (S) 296 <i>Salmonella typhimurium</i> d-ribose-binding periplasmic protein precursor	31.6% 304 aa
<i>f307</i>	307 aa	YNGB_CLOPE (S) 279 <i>Clostridium perfringens</i> hypothetical ORF product	25.9% 264 aa
<i>o149</i>	149 aa	Similar to galactose-6-phosphate isomerase from several sources:	
		D43258 (N) 171 <i>Streptococcus mutans</i> galactose-6-phosphate isomerase subunit	40.9% 148 aa
		LACB_STRMU (S) 171 <i>Streptococcus mutans</i> galactose-6-phosphate isomerase subunit	33.1% 141 aa
		LACB_STAAU (S) 171 <i>Staphylococcus aureus</i> galactose-6-phosphate isomerase	31.7% 141 aa
		LACB_LACLA (S) 171 <i>Lactococcus lactis</i> galactose-6-phosphate isomerase	35.9% 141 aa
<i>o742</i>	742 aa	PS00017 (P) ATP_GTP_A [ATP/GTP-binding site motif A (P-loop)]	100% match
<i>f577</i>	577 aa	STYPMTRAB_1 (G) 547 <i>Salmonella typhimurium</i> hypothetical ORF product in <i>pmr</i> region	82.4% 547 aa
<i>f326</i>	326 aa	Similar to antiporter proteins in a permease family that includes ArcD, CadB, GerAII and PotE	
		CADB_ECOLI (S) 444 <i>Escherichia coli</i> putative cadaverine/lysine antiporter	44.6% 322 aa
		POTE_ECOLI (S) 439 <i>Escherichia coli</i> putrescine-ornithine antiporter	36.2% 320 aa
		ARCD_PSEAE (S) 482 <i>Pseudomonas aeruginosa</i> putative arginine/ornithine antiporter	28.2% 324 aa
<i>f239</i>	239 aa	BACGDH1_1 228 (G) <i>Bacillus megaterium</i> hypothetical ORF product in glucose dehydrogenase region	34.8% 224 aa
<i>f485</i>	485 aa	LLU05215_1 463 (G) <i>Lactococcus lactis</i> ML3 di-tripeptide transporter	25.0% 467 aa
<i>f177</i>	177 aa	PS00213 (P) LIPOCALIN [Lipocalin signature]	100% match
		VCRVC_3 (G) 171 <i>Vibrio cholerae</i> Vlp antigen (vibrio lipoprotein) in RVC repeat region	51.0% 150 aa
<i>f337</i>	337 aa	PS00017 (P) ATP_GTP_A [ATP/GTP-binding site motif A (P-loop)]	100% match
<i>f379</i>	379 aa	PS00198 (P) 4FE4S_FERREDOXIN [4Fe-4S ferredoxins, iron-sulfur binding region signature]	100% match
		YGL5_BACST (S) 216 <i>Bacillus stearothermophilus</i> hypothetical ORF product in <i>gldA</i> region	33.7% 199 aa
<i>o153</i>	153 aa	PS00198 (P) 4FE4S_FERREDOXIN [4Fe-4S ferredoxins, iron-sulfur binding region signature]	100% match
<i>o243</i>	243 aa	BACGLUSYN_7 (G) 144 <i>Bacillus subtilis</i> hypothetical ORF product in <i>gltX-cysES</i> region	32.4% 140 aa
		TSR_STRAZ (S) 269 <i>Streptomyces azureus</i> rRNA methyltransferase	27.0% 145 aa
		ECOSPOT_3 (G) 229 <i>Escherichia coli</i> hypothetical ORF product in <i>spoT</i> region	28.8% 157 aa
<i>f251</i>	251 aa	PS00894 (P) HTH_DEOR_FAMILY [Bacterial regulatory proteins, deoR family signature]	100% match
		LACR_LACLA (S) 255 <i>Lactococcus lactis</i> lactose phosphotransferase system repressor protein	27.1% 254 aa
		LACR_STRMU (S) 251 <i>Streptococcus mutans</i> lactose phosphotransferase system repressor protein	26.9% 245 aa
		LACR_STAAU (S) 251 <i>Staphylococcus aureus</i> lactose phosphotransferase system repressor protein	21.9% 240 aa
		ACCR_AGRTU (S) 258 <i>Agrobacterium tumefaciens</i> AccR, transcriptional repressor protein	22.0% 246 aa
<i>o158</i>	158 aa	PS00387 (P) PPASE [Inorganic pyrophosphatase signature]	100% match
		PTYA_ECOLI (S) 147 <i>Escherichia coli</i> cryptic mannitol-specific permease, PTS system enzyme IIA	35.5% 149 aa
		PTMA_ECOLI (S) 637 <i>Escherichia coli</i> MtlA, mannitol-specific phosphotransferase enzyme II, mannitol permease	31.5% 138 aa
<i>o228</i>	228 aa	ARAD_ECOLI (S) 231 <i>Escherichia coli</i> AraD, L-ribulose-5-phosphate 4-epimerase	59.9% 230 aa
		ARAD_SALTY (S) 248 <i>Salmonella typhimurium</i> AraD, L-ribulose-5-phosphate 4-epimerase	62.6% 196 aa
<i>o259</i>	259 aa	MIP_LEGPN (S) 233 <i>Legionella pneumophila</i> Mip, outer membrane protein precursor, macrophage infectivity potentiator, PPIase	41.1% 207 aa
		MIP_LEGMI (S) 243 <i>Legionella micdadei</i> Mip, outer membrane protein precursor, macrophage infectivity potentiator, PPIase	38.2% 226 aa
		CTMIPB_1 (G) 243 <i>Chlamydia trachomatis</i> (serovar B) Mip-like protein	32.2% 174 aa
		CTMIPE_1 (G) 243 <i>Chlamydia trachomatis</i> (serovar E) Mip-like protein	32.2% 174 aa
<i>o470</i>	470 aa	STYCARABA_3 (G) 461 <i>Salmonella typhimurium</i> putative transport protein	34.8% 452 aa
		AROP_ECOLI (S) 456 <i>Escherichia coli</i> AroP, aromatic amino acid transport protein	39.2% 455 aa
		PHEP_ECOLI (S) 458 <i>Escherichia coli</i> PheP, phenylalanine-specific permease	37.2% 464 aa
		YIFK_ECOLI (S) 461 <i>Escherichia coli</i> O461, putative amino acid transport protein	37.7% 458 aa

Table 2. continued

<i>o577</i>	577 aa	PS00343 (P) GRAM_POS_ANCHORING [Gram-positive cocci surface proteins 'anchoring' hexapeptide]	100% match	
<i>o492</i>	492 aa	PS00213 (P) LIPOCALIN [Lipocalin signature]	100% match	
<i>o116</i>	116 aa	R10REP_6 (G) 133 <i>Escherichia coli</i> plasmid R100 hypothetical ORF product in <i>repA</i> region	43.1%	114 aa
<i>o417</i>	417 aa	PS00017 (P) ATP_GTP_A [ATP/GTP-binding site motif A (P-loop)]	100% match	
		RBSA_ECOLI (S) 501 <i>Escherichia coli</i> RbsA, high affinity ribose transport protein	36.0%	358 aa
		ARAG_ECOLI (S) 504 <i>Escherichia coli</i> AraG, high affinity L-arabinose transport protein	35.6%	361 aa
		YJCW_ECOLI (S) 502 <i>Escherichia coli</i> (F502, partial in ECOUW89) hypothetical ABC transporter in <i>fdhF</i> region	33.2%	364 aa
		MGLA_ECOLI (S) 506 <i>Escherichia coli</i> MglA, galactoside-binding protein	34.4%	343 aa
<i>o90</i>	90 aa	MGLA_ECOLI (S) 506 <i>Escherichia coli</i> MglA, galactoside-binding protein	52.2%	67 aa
		RBSA_ECOLI (S) 501 <i>Escherichia coli</i> RbsA, high affinity ribose transport protein	44.7%	85 aa
<i>o341</i>	341 aa	RBSC_ECOLI (S) 321 <i>Escherichia coli</i> RbsC, high affinity ribose transport protein	35.3%	326 aa
		ARAH_ECOLI (S) 329 <i>Escherichia coli</i> AraH, high affinity L-arabinose transport protein	35.5%	258 aa
		YJCV_ECOLI (S) 326 <i>Escherichia coli</i> (F326 in ECOUW89) hypothetical ORF product, <i>fdhF</i> region	29.8%	269 aa
<i>o323</i>	323 aa	RBSC_ECOLI (S) 321 <i>Escherichia coli</i> RbsC, high affinity ribose transport protein	34.5%	283 aa
		ARAH_ECOLI (S) 329 <i>Escherichia coli</i> AraH, high affinity L-arabinose transport protein	29.7%	300 aa
		MGLC_ECOLI (S) 336 <i>Escherichia coli</i> MglC, galactose transport protein	29.1%	314 aa
		YJCV_ECOLI (S) 326 <i>Escherichia coli</i> (F326 in ECOUW89) hypothetical ORF product, <i>fdhF</i> region	30.7%	282 aa
<i>o457</i>	457 aa	MURC_ECOLI (S) 491 <i>Escherichia coli</i> MurC, UDP-N-acetylmuranoyl-L-alanine synthetase	25.8%	400 aa
<i>f154</i>	154 aa	Y559_BPT4 (S) 156 bacteriophage T4 hypothetical ORF product in gp55- <i>sunY</i> region	46.4%	148 aa
<i>mgIA</i>	898 aa	STU07843_2 (G) <i>Salmonella typhimurium</i> MgtA, Mg++ transport P-type ATPase	88.4%	896 aa
<i>f261</i>	261 aa	PS00061 (P) ADH_SHORT [Short-chain alcohol dehydrogenase family signature]	2 x 100% matches	
<i>o138</i>	138 aa	STMIAE_2 138 (G) <i>Salmonella typhimurium</i> hypothetical ORF product in <i>miaE</i> , <i>argI</i> region	92.0%	138 aa
<i>f97</i>	97 aa	STMIAE_5 167 (G) <i>Salmonella typhimurium</i> hypothetical ORF product in <i>miaE</i> , <i>argI</i> region	81.4%	97 aa
<i>o204</i>	204 aa	STMIAE_6 268 (G) <i>Salmonella typhimurium</i> hypothetical ORF product in <i>miaE</i> , <i>argI</i> region	34.4%	120 aa
		VP4_ROTTHD (S) 775 <i>Homo sapiens</i> rotavirus outer capsid protein vp4	26.8%	81 aa
<i>f500</i>		PS00017 (P) ATP_GTP_A [ATP/GTP-binding site motif A (P-loop)]	100% match	
<i>f332</i>	332 aa	Similar to several repressor proteins: CYTR_ECOLI (S) 341 <i>Escherichia coli</i> CytR, repressor protein	25.1%	329 aa
		CLOREPA_1 (G) 332 <i>Clostridium acetobutylicum</i> RepA, regulatory protein	25.6%	324 aa
		GALR_HAEIN (S) 332 <i>Haemophilus influenzae</i> GalR, galactose operon repressor protein	25.3%	319 aa
		GALR_ECOLI (S) 343 <i>Escherichia coli</i> GalR, galactose operon repressor protein	23.8%	304 aa
		PURR_ECOLI (S) 341 <i>Escherichia coli</i> PurR, purine nucleotide synthesis repressor protein	22.7%	332 aa
		BACDEGARP_1 (G) 337 <i>Bacillus subtilis</i> DegA, repressor-like protein	23.0%	300 aa
<i>f439</i>	439 aa	GNTP_BACSU (S) 448 <i>Bacillus subtilis</i> GntP, gluconate permease	49.1%	438 aa
<i>f254</i>	254 aa	PS00061 (P) ADH_SHORT [Short-chain alcohol dehydrogenase family signature]	100% match	
		ECKDG_3 (G) 253 <i>Erwinia chrysanthemi</i> KdkD, 2-keto-3-deoxygluconate oxydoreductase	41.8%	248 aa
		FABG_ECOLI (S) 244 <i>Escherichia coli</i> FabG, 3-ketoacyl-acyl carrier protein reductase	33.5%	247 aa
		BSGENR_83 (G) 255 <i>Bacillus subtilis</i> hypothetical ORF product	32.9%	252 aa
		DHG_BACSU (S) 260 <i>Bacillus subtilis</i> Gdh, glucose 1-dehydrogenase	33.3%	246 aa
		NODG_AZOBR (S) 246 <i>Azospirillum brasilense</i> NodG, nodulation protein, ribitol dehydrogenase family	34.7%	245 aa
		ATBORED_1 (G) 319 <i>Arabidopsis thaliana</i> , 3-oxoacyl-[acyl-carrier protein] reductase, ribitol dehydrogenase family	33.3%	243 aa
		YRTP_BACSU (S) 238 <i>Bacillus subtilis</i> hypothetical ORF product in <i>rtp</i> region	36.1%	194 aa
		BACENTA_1 (G) 196 <i>Bacillus subtilis</i> EntA, 2,3-dihydro-2,3-dihydroxybenzoate dehydrogenase	35.7%	140 aa
<i>f343</i>	343 aa	Similar to several sorbitol dehydrogenases: DHSD_HUMAN (S) 355 <i>Homo sapiens</i> sorbitol dehydrogenase	32.7%	307 aa
		DHSD_SHEEP (S) 354 <i>Ovis aries</i> sorbitol dehydrogenase	32.2%	321 aa
		BACGUTB_1 (G) 353 <i>Bacillus subtilis</i> GutB, sorbitol dehydrogenase	29.1%	328 aa
<i>gntV</i>	187 aa	PS00017 (P) ATP_GTP_A [ATP/GTP-binding site motif A (P-loop)]	100% match	
<i>f353</i>	353 aa	PS00059 (P) ADH_ZINC [Zinc-containing alcohol dehydrogenases signature]	100% match	
		PS00190 (P) CYTOCHROME_C [Cytochrome c family heme-binding site signature]	100% match	
		ADH_MYCBO (S) 346 <i>Mycobacterium bovis</i> Adh, NADP-dependent alcohol dehydrogenase	36.2%	343 aa
		ATELI32_1 (G) 359 <i>Arabidopsis thaliana</i> Eli3-2, plant defense protein	35.1%	337 aa

Table 2. continued

		PACINALDA_1 (G) 357 <i>Picea abies</i> cinnamyl alcohol dehydrogenase	35.4%	341 aa
		CADH_MEDSA (S) 358 <i>Medicago sativa</i> cinnamyl-alcohol dehydrogenase	33.7%	337 aa
		YCZ5_YEAST (S) 361 <i>Saccharomyces cerevisiae</i> hypothetical ORF product	32.5%	330 aa
o396	396 aa	PS00018 (P) EF_HAND [EF-hand calcium-binding domain]	100%	match
		VINT_BPP4 (S) 440 bacteriophage P4 integrase	58.5%	386 aa
		SF6INTOAC_1 (G) 385 <i>Shigella flexneri</i> bacteriophage Sf6 integrase	31.1%	376 aa
o109	109 aa	ATACH5_14 (G) 104 <i>Agrobacterium tumefaciens</i> Ti plasmid ORF product	42.9%	28 aa
o425	425 aa	<i>Escherichia coli</i> NanT, sialic acid permease (personal communication, E. Vimr)	30.0%	405 aa
		PPU10895_1 (G) 449 <i>Pseudomonas putida</i> PcaK, 4-hydroxybenzoate transporter	26.5%	195 aa
		ACCPAOP_5 (G) 422 <i>Acinetobacter calcoaceticus</i> PcaORF, putative transport protein	25.0%	190 aa
		PROP_ECOLI (S) 500 <i>Escherichia coli</i> ProP, proline/betaine transporter PPII	23.0%	396 aa
		PMTETRH_2 (G) 400 <i>Pasteurella multocida</i> TetA, tetracycline resistance determinant	20.4%	385 aa
f388	388 aa	Gntp_BACSU (S) 448 <i>Bacillus subtilis</i> GntP, gluconate permease	57.1%	355 aa
f655	655 aa	ILVD_ECOLI (S) 605 <i>Escherichia coli</i> IlvD, dihydroxy-acid dehydratase	27.8%	535 aa
		EDD_ZYMMO (S) 583 <i>Zymomonas mobilis</i> Edd, 6-phosphogluconate dehydratase	28.7%	505 aa
f319	319 aa	DAPA_BACSU (S) 290 <i>Bacillus subtilis</i> DapA, dihydrodipicolinate synthase	29.9%	282 aa
		DAPA_ECOLI (S) 292 <i>Escherichia coli</i> DapA, dihydrodipicolinate synthase	28.1%	297 aa
		BLDAPAB_3 (G) 301 <i>Brevibacterium lactofermentum</i> DapA, dihydrodipicolinate synthase	29.0%	309 aa
		RHMMOSA_2 (G) 333 <i>Rhizobium meliloti</i> MosA protein	30.8%	275 aa
		DAPA_CORGL (S) 301 <i>Corynebacterium glutamicum</i> DapA, dihydrodipicolinate synthase	29.0%	309 aa
f262	262 aa	ECKDGR_1 (G) 306 <i>Erwinia chrysanthemi</i> KdgR repressor protein	31.6%	226 aa
		ICLR_ECOLI (S) 274 <i>Escherichia coli</i> IclR, acetate operon repressor protein	25.8%	248 aa
f260	260 aa	ACCR_AGRU (S) 258 <i>Agrobacterium tumefaciens</i> plasmid pTIC58 AccR, transcriptional repressor	28.3%	239 aa
		DEOR_ECOLI (S) 252 <i>Escherichia coli</i> DeoR, deoxyribose operon repressor protein	23.3%	251 aa
f210	210 aa	YJCU_ECOLI (S) 231 <i>Escherichia coli</i> (F231 in ECOUW89) hypothetical ORF product in <i>fdhF</i> region	30.6%	205 aa
		YHFD_ECOLI (S) 221 <i>Escherichia coli</i> hypothetical ORF product in <i>trpS-dam</i> region	28.0%	208 aa
		PG1CFXXYEP_3 (G) 241 <i>Alcaligenes eutrophus</i> plasmid pHG1 CfxEp, D-ribulose-5-phosphate 3 epimerase	29.0%	206 aa
f143	143 aa	PTYA_ECOLI (S) 147 <i>Escherichia coli</i> cryptic mannitol-specific permease, PTS system enzyme IIA	34.6%	135 aa
f437	437 aa	PS00075 (P) DHFR [Dihydrofolate reductase signature]	100%	match
f383	383 aa	YIII_ECOLI (S) 356 <i>Escherichia coli</i> (F356 in ECOUW87) hypothetical ORF product in <i>fdhD</i> region	29.3%	346 aa
f477	477 aa	Gntp_BACSU (S) 448 <i>Bacillus subtilis</i> gluconate permease	36.5%	453 aa
uxuB	486 aa	ECOHU47_61 (G) 488 <i>Escherichia coli</i> YeiQ hypothetical ORF product in <i>fruB</i> region	55.6%	443 aa
uxuR	257 aa	PS00043 (P) HTH_GNTR_FAMILY [Bacterial regulatory proteins, gntR family signature]	100%	match
f257	257 aa	AFHGDC_2 (G) 260 <i>Acidaminococcus fermentans</i> HgdC, activator of (R)-hydroxyglutaryl-CoA dehydratase	42.2%	250 aa
f410	410 aa	PS00216 (P) SUGAR_TRANSPORT_1 [Sugar transport proteins signature 1]	100%	match
		CMLA_PSEAE (S) 419 <i>Pseudomonas aeruginosa</i> CmlA, chloramphenicol resistance protein	25.4%	318 aa
		YIDY_ECOLI (S) 391 <i>Escherichia coli</i> (O392 in ECOUW82) hypothetical ORF product in <i>maB</i> region	23.8%	386 aa
o103	103 aa	YHGA_ECOLI (S) 292 <i>Escherichia coli</i> hypothetical ORF product in <i>feoAB</i> region	59.2%	103 aa
		YADD_ECOLI (S) 283 <i>Escherichia coli</i> hypothetical ORF product in <i>panBCD</i> region	59.5%	84 aa
f470	470 aa	BSSRFAP_6 (G) 444 <i>Bacillus subtilis</i> hypothetical ORF product in <i>srfA-sfp</i> region	25.0%	382 aa
f67	67 aa	YBDD_ECOLI (S) 65 <i>Escherichia coli</i> hypothetical ORF product in <i>cstA</i> region	71.0%	67 aa
f721	721 aa	CSTA_ECOLI (S) 562 <i>Escherichia coli</i> CstA, carbon starvation protein	66.0%	556 aa
		YBDC_ECOLI (S) 138 <i>Escherichia coli</i> hypothetical ORF product in <i>cstA</i> region	42.0%	168 aa
f192	192 aa	YIDT_ECOLI (S) 445 <i>Escherichia coli</i> (F445 in ECOUW82), hypothetical ORF product in <i>gyrB</i> region	29.5%	180 aa
f261	261 aa	YIDT_ECOLI (S) 445 <i>Escherichia coli</i> (F445 in ECOUW82), hypothetical ORF product in <i>gyrB</i> region	29.0%	206 aa
o345	345 aa	PS00272 (P) SNAKE_TOXIN [Snake toxins signature]	100%	match
		TDH_ECOLI (S) 341 <i>Escherichia coli</i> Tdh, threonine 3-dehydrogenase	28.4%	329 aa
f343	343 aa	PS00092 (P) N6_MTASE [N-6 Adenine-specific DNA methylases signature]	100%	match
		BACRPLL_2 (G) 201 <i>Bacillus subtilis</i> RpoL, ribosomal protein L7/12	26.0%	202 aa

Table 2. continued

<i>o211</i>	211 aa	PS00343 (P) GRAM_POS_ANCHORING [Gram-positive cocci surface proteins 'anchoring' hexapeptide]	100% match
<i>f287</i>	287 aa	PS00198 (P) 4FE4S_FERREDOXIN [4Fe-4S ferredoxins, iron-sulfur binding region signature]	2 x 100% matches
		PS0019 (P) CYTOCHROME_C [Cytochrome c family heme-binding site signature]	100% match
<i>f579</i>	579 aa	PS00211 (P) ABC_TRANSPORTER [ABC transporters family signature]	100% match
		PS00017 (P) ATP_GTP_A [ATP/GTP-binding site motif A (P-loop)]	2 x 100% matches
		CHVD_AGRTU (S) 286 <i>Agrobacterium tumefaciens</i> ChvD protein (fragment)	56.0% 283 aa
<i>f173</i>	173 aa	PS00216 (P) SUGAR_TRANSPORT_1 [Sugar transport proteins signature 1]	100% match
		ENTTRPRA_3 (G) 67 <i>Enterobacter cloacae</i> hypothetical ORF product in <i>trpR</i> region	84.8% 66 aa
		ENTTRPR_2 (G) 54 <i>Enterobacter aerogenes</i> hypothetical ORF product in <i>trpR</i> region	88.7% 53 aa
<i>o215</i>	215 aa	PS00175 (P) PG_MUTASE [Phosphoglycerate mutase family phosphohistidine signature]	100% match

Database sequences were aligned with the predicted proteins, using DNASTAR's Align for the Macintosh with a gap penalty of 4 and a gap length penalty of 12; the score is the percent amino acid identity and the length of the alignment. The locus name, source database [(S) SwissProt release 27, (N) NBRF-PIR release 40, (G) translated GenBank release 82] and length in amino acids is listed for each database entry. Sequences were also compared with the PROSITE database [(P), release 11] using MacPattern (46). The products of previously sequenced and characterized genes are not listed in this table.

55.4% similar to the *E.coli* NrdD over 614 aa), separated by gene 55.10. Interestingly, the T4 *nrdD* gene contains a self-splicing group I intron, and it has been suggested that the gene was introduced into T4 via horizontal transfer (26). If such an event took place, gene 55.9 was probably acquired at the same time; the function of T4 gene 55.10, for which there is no corresponding ORF in *E.coli*, is unknown.

ORFs *o417*, *o90*, *o341* and *o323* (141 888–145 407) may be components of an ATP-binding sugar transport (PTS) operon. Individual ORFs show similarities in the 30–40% range to members of the *rbs*, *ara* and *mgl* operons, and the next gene downstream is *fbp* (fructose biphosphatase; also called *fdp*). The distribution of similarities and the codon usage statistics of ORFs *o417* and *o90* suggest these may really be one ORF but we were unable to find a frameshift error.

Other examples of suggested gene function include *o149* at 4279 (galactose-6-phosphate isomerase), *f326* at 26879 (antiporter), *f251* at 108530 (PTS system repressor), *f254* at 183442, *f343* at 184210 and *f353* at 186025 (dehydrogenases).

Genes mapped but not found

Twenty-one genes map to this region but could not be identified with confidence although matches to entries in the protein databases noted in Table 2 may suggest candidates for some: *gltR*, *aidC*, *pheR*, *dgkR*, *ssaH*, *poxA*, *abs*, *thdC*, *calA*, *corB*, *sbaA*, *prp*, *cup*, *oppE*, *dadQ*, *fhuF*, *mdoB*, *murH*, *radA*, *thdD* and *ilvR*.

Other features

Although few insertion sequences have been found so far in MG1655, four occur in this segment: IS2, IS4, IS30 and IS1F. The attachment site for bacteriophage P4 (*attP4*) is found at 187 305–187 324, within the tRNA gene *leuX*. The adjacent ORF, *o396* encodes a product 58.5% similar to the integrase of P4, and 31.1% similar to the integrase of the Shigella bacteriophage Sf6. Whether the putative O396 has recombinase activity is not known. It may simply be a degenerate relic of imperfect excision by P4. In addition to *leuX* already mentioned, there are seven other transfer RNAs in this segment: *leuVPQ*, *pheU* and *glyVXY*.

Table 3. Genes encoding predicted signal peptides

Known genes	Reference	ORFs	Coordinate
<i>fecB</i>	(47)	<i>f311</i>	2036
<i>fecA</i>	(47)	<i>f177</i>	68 017
<i>fimA</i>	(48)	<i>f1107</i>	76 875
<i>fimF</i>	(48)	<i>o212</i>	103 215
<i>fimG</i>	(48)	<i>f142</i>	107 273
<i>fimH</i>	(48)	<i>f91</i>	115 348
<i>fimC</i>	(49)	<i>f184^a</i>	129 539
<i>fimI</i>	(50)	<i>o577</i>	133 213
<i>fimD</i>	(51)	<i>o492</i>	134 943
<i>ampC</i>	(52)	<i>o318^a</i>	140 794
<i>creA</i>	(53)	<i>o164</i>	165 058
<i>creC</i>	(53)	<i>o204</i>	171 144
<i>osmY^a</i>	(54)	<i>f404</i>	228 491
<i>cycZ</i>	(55)	<i>f261</i>	284 770
<i>phnD</i>	(40)	<i>f165</i>	290 527
<i>phnE</i>	(40)		
<i>cpdB^a</i>	(56)		
<i>smp</i>	(57)		
<i>amiB</i>	(58)		
<i>slt</i>	(58)		
<i>frdD^a</i>	(60)		
<i>frdC</i>	(60)		

^aGenes also found in the Church–Link database to have N-termini corresponding to cleavage of the putative signal peptide.

There are two areas in this segment which might be grey holes, that is, regions containing no features discernable by current search methods (1). They span the regions 224 663–225 737 between *f181* and *o338*, and 230 538–231 598 between *f241* and *fimB*, and are 1075 and 1242 bases long, respectively. A careful

Figure 1.

Figure 1. continued

Figure 1. Map of the finished sequence and its features, proceeding left to right in 16 tiers. The top line shows restriction sites for 10 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 sigmas 28, 54 and 32, respectively; unlabelled promoters are regulated by $\sigma 70$. Other sequence determinations spanned by this segment are indicated, with their database accession numbers. The repeat sequences near 234 kb, marked 'fim invert', denote the invertible element controlling FimA expression; in this sequence the element is in the OFF orientation. The annotated sequence is in GenBank (accession no. U14003).

screening of these regions for ORFs revealed some possible very short candidates, but the codon usage patterns were not sufficiently convincing to define these for annotation.

Chi site and translation orientation. A survey of chi sites within the coding regions showed that of the 48 sites found, 26 are oriented with both replication and translation, 19 are oriented with either replication or translation, and only three are oriented against both. This correlation is generally in agreement with that observed previously (2–5), though the correlation between translation and replication orientation has not been maintained (2). There are 319 genes in this segment which account for 86.7% of the sequence. Assessment of gene orientation with respect to the direction of replication reveals 44% of expressed DNA is translated in the same direction as replication and 56% is expressed in the opposing direction.

Signal peptides. A computer search for aa sequences that resemble signal peptides was carried out by a specially written program using the weight-matrix method of von Heijne (27). The results are shown in Table 3. The 22 known genes listed were previously demonstrated to be membrane bound or processed soluble proteins. Of the ORFs listed, two (O318 and F184) were also found in the Church-Link database of experimentally determined

N-terminal sequences. [This database is available via World Wide Web at [ftp://rascal.med.harvard.edu/churchlab/protein/](http://rascal.med.harvard.edu/churchlab/protein/) and has also been integrated into SwissProt (28).] The database indicated both of these putative ORF products to have processed mature N-termini at residue 22 corresponding to cleavage at the end of the putative signal peptide. Four of the ORFs listed showed similarity to other proteins or motifs in the sequence databases (Table 2) which concur with their putative signal sequences. The putative product of *f311* is similar to RbsB proteins of *S.typhimurium* and *E.coli*, the precursors of ribose-binding periplasmic proteins. The putative ORF products F177 and O492 contain 100% matches to the lipocalin signature of eukaryotic proteins that transport small hydrophobic molecules. The predicted product of *o577* contains a 100% match to the conserved anchoring hexapeptide in surface proteins of Gram-positive cocci.

The oligonucleotide CTAG is rare in the *E.coli* genome

The occurrence of di-, tri- and tetranucleotides in the genomic DNA of several species has been studied (29,30) and shown to be non-uniform. In previous surveys, non-contiguous *E.coli* sequences were analyzed. To create a large contiguous genomic sequence for such analyses we joined the published megabase of sequence from this group and the published 0–4.1 minute

segment (6,7). The resulting sequence of 1.6 Mb was used to examine the distribution in coding and non-coding sequence of the tetramer CTAG, known to be rare in *E. coli* (31–33).

Table 4. Frequency of rare tetranucleotides in 1.6 Mb of *E. coli*

		Sites	kb	Average spacing (kb/sites)
CTAG	Total	317	1629.0	5.1
	In genes	190	1362.9	7.1
	In rRNA	40	30.2	0.75
	Intergenic	87	225.6	2.6
	Promoters	30	57.0	1.9
TTAA	Total	7188	1629.0	0.23
	In genes	5270	1362.9	0.26
	Intergenic	1918	266.1	0.12

Three hypotheses have been proposed to explain the rarity of CTAG. (i) Codon bias due to amber stops (TAG) in both orientations (29). (ii) Very Short Patch Repair, the system which corrects mismatches, could remove CTAG from the genome if its specificity is poor (34–36). Normally, a methylated deoxycytosine in CCAGG may be deaminated to thymine, producing CTAGG with a T:G mismatch, which the VSPR system will correct to C:G. If this mechanism operates erroneously on a natural CTAG not mismatched, these sequences could be progressively removed. (iii) The ability of DNA to bend or 'kink' at CTAG may serve a function when it occurs in a repressor binding site (29,30), presumably to mediate the biological activity of the repressor protein, but be disadvantageous elsewhere.

Does our survey of CTAG sites in the large contiguous sequence distinguish among these hypotheses? Not entirely, but we do find some interesting arguments for codon bias.

The sequence CTAG would be expected to occur at random every 256 nt. Instead we found 317 sites, spaced on average at 5.1 kb, or 20-fold less frequent than expected for random occurrence (Table 4). The distribution was not random, however. Table 4 shows the spacing of CTAG sites in coding regions, non-coding regions, RNA genes and promoters. The low frequency of sites is most striking in protein-coding sequence, where 190 sites were found in 1.36 Mb, with an average spacing of 7.1 kb, or 28-fold fewer than expected. By comparison, the frequency in intergenic regions is only 10-fold lower than expected, and the 40 sites in rRNA genes are much closer to the expected frequency being only 3-fold lower (Table 4).

This distribution pattern suggests that the coding capacity of the tetramer might be a biasing factor, in particular the amber (TAG) codons in both orientations in CTAG. However, as noted by Burge (29), the tetramer TTAA which similarly encodes stops in both orientations, is found at the frequency expected for random occurrence. We also found that the distribution of TTAA in coding and non-coding DNA was biased towards intergenic regions, but less so than for CTAG: a total of 7188 sites were spaced on average at intervals of 259 bases in genes and 118 bases in intergenic regions (Table 4).

Table 5. (A) Reading frame of 190 CTAG tetramers in the 1.6 Mb sequence

Number (%)	Frame	Codon	Amino acid	Codon usage %
10 (5.3)	C TAG	TAG	amber	8
49 (25.8)	CT AG			
12		ACT	thr	16
25		GCT	ala	15
9		CCT	pro	14
3		TCT	ser	15
4		AGA	arg	3
0		AGG	arg	2
34		AGC	ser	27
11		AGT	ser	13
13 (68.9)	CTA G	CTA	leu	3
1				

Table 5. (B) Codon usage for leucine, arginine and serine in the 1.6 Mb sequence

Leu	%	Arg	%	Ser	%	Stop codons	%
CUA	3	AGA	3	UCA	11	UAG	8
UUA	11	CGA	5	UCC	16	UAA	28
CUC	10	CGC	41	AGC	27	UGA	63
CUG	53	AGG	2	UGC	15		
UUG	12	CGG	9	UCU	15		
CUU	9	CGU	39	AGU	13		

Of the 1432 genes in this sequence 167 contain at least one CTAG site, 190 sites in total. To investigate further the possible coding bias, we examined these instances of sites within genes for the reading frame used (Table 5A). In 10 cases, the amber codon was in frame. UAG is the least frequently used stop codon in this sequence (Table 5B). In 49 cases, the frame was XCT AGX, encoding (serine, proline, threonine or alanine) and (serine or arginine). The occurrence of each of these possible codons is shown in Table 5A. In general, the usage of these codons within CTAG does not differ significantly from the overall usage. In striking contrast, however, 132 cases (70%), were in the frame CTA:GXX. CTA is the rarest leucine codon, used for only 3% of the leucines in the segment.

In all three frames, then, there is an apparent coding constraint. One of the codons in each possible frame is normally very rarely used or is a stop. If rare codons are typically translated with poor efficiency, CTAG may have been selected against in evolution. There is a suggestion that rare codons may have a role in growth control through a variety of *E. coli* genes where CTAG persists (37) for the the rare arginine codons AGA and AGG. These arguments might account for the predominance of the third reading frame CTA, the rarest leucine codon, if this leucine codon were nevertheless easier to translate than the rare arginine codon, perhaps due to differences in the availability of the corresponding tRNAs.

Analysis of CTAG in promoter regions in the 1.06 Mb of sequence that is annotated, showed that the tetramer was found more frequently in these areas than in coding or intergenic regions (Table 4). The sequence was searched from 50 bases upstream to

20 bases downstream of the -10 and -35 elements, a total span of ~100 bases in each case. Twenty-seven promoter regions contained CTAG, three of them twice, giving an average spacing of only 1.9 kb for these regions. Thus CTAG occurs 3.7-fold more frequently in promoter regions than within genes, a small effect but consistent with the idea that these sites might function in gene expression control.

This analysis does not eliminate any of the hypotheses put forward to explain CTAG rarity, and indeed allows that all three might operate in some proportion, (i) and (iii) acting in coding and non-coding regions, respectively, and (ii) more generally. Perhaps VSPR is the mechanism for generating mutations which are then selected by pressures for codon bias and promoter activity, so that all three make essential contributions.

Differences

phn operon. In K-12 strains this operon is cryptic due to a frameshift in *phnE* (38). We have confirmed that in MG1655 *phnE* is split into two ORFs in comparison with the same gene in *E. coli* B, by an extra copy of an eight base repeat, making three in K-12 and two in B (39). The gene previously termed *phnQ* is not essential for phosphonate utilisation (40,41) and is probably misnamed.

hsdR. Our sequence differs from the previously published sequence of *hsdR* in two respects: an internal in-and-out of frame difference alters a stretch of 69 aa (residues 629–697), and a frameshift changes both the sequence and the length beyond aa 1083 (for a net length of 1090 aa in X06545, as compared with 1188 aa in the sequence described here). Examination of both sequences using GeneMark indicates potential frameshifts in X06545, at the positions where the sequences differ. In addition, the predicted protein product of the sequence presented here retains all of the conserved motifs described by Murray *et al.* (42) for type I restriction enzyme R subunits, and displays additional similarities to *EcoA* and *EcoE* HsdR sequences in the region where we differ from the previous *EcoK* HsdR sequence (alignments not shown).

sugEF. The published sequence of *sugE* (X69949), derived from *E. coli* K-12 strain CS520, contains an unusual arrangement of 'nested' genes. There is evidence that both are transcribed and translated (43). We found that MG1655 is clearly different due to an insertion of three extra bases creating a stop codon. The larger gene of the 'nested' pair is thus 17 codons shorter at the 5' end than that reported. The smaller coding sequence begins downstream of the difference and is presumably unaffected. The biological significance, if any, of this strain difference is unknown.

100/0 minute point

How should the 100/0 minute point of the physical map be defined within the sequence? The original definition was purely genetic, made by measuring time of entry of marker genes in conjugation experiments. It would be useful to make a precise physical assignment now that the sequence of the area is completed.

The first gene on the genetic map is *thrA* (10). One logical possibility for the zero point would be the insertion site of HfrH, the true point of origin for the original interrupted mating experiments in which *thrA* had the shortest time of entry. Another possibility is to arbitrarily select a site upstream of *thrA*, for

example a restriction site such as the *HindIII* site in *o228* or a site between *o228* and *thrL* in non-coding DNA. In their sequence determination of the 0–2.4 minute region, Yura *et al.* (6) chose the first base of *thrA*. The integrated map and sequence Ecomap compiled by K. Rudd starts at the *HindIII* site in the end of *o228*. Our own sequence overlaps with pre-existing determinations of the *thr* region and contains no annotation to define this point yet; the first A of the *HindIII* site as base 1 seems to us to be the most practical solution at this time.

Update: correction to ECOUW93 after submission to GenBank

The following corrections will be made to ECOUW93, changing *f72* at 218 594–218 812 to *f143*: delete C at position 218 605, and add G between positions 218 574 and 218 575. No other ORFs or features are affected. Whereas the putative *f72* product was 38.1% similar to an *E. coli* putative phosphotransferase III, that of *f143* shows a similar match (34.6%) over its whole length (Table 2), which is much closer to the average of 150 aa for other phosphotransferases, and includes the conserved motifs of this protein family. In addition, *f750* (287 982–290 234) is identified as *mdoB* by comparison of the restriction map (44) with ECOUW93 in this region.

ACKNOWLEDGEMENTS

This is paper 3416 from the Laboratory of Genetics, supported by award HG000301 from the NIH Human Genome Project. We thank M. Kröger and K. Rudd for their continuing help with databases, E. Vmir for unpublished information about *nanT*, K. Rudd and T. Conway for unpublished information on *gnt* genes, A. Bairoch for pointing out the membrane-spanning domains in the putative F721, J. Reizer for pointing to the errors in *f72* and K. Zahn and S. Karlin for helpful discussions. For software development and assistance we thank M. Borodovsky, G. Bouriakov, D. Argentar and the programming staff of DNASTAR. Excellent technical support was provided by R. Clark, K. Kadner, H. Kirkpatrick, M. Maguire, D. Rose, N. Shi, S. Subramanian and 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: S. Abdul-Jabbar, E. Ahmad, H. Cheng, W. Davis, D. Duescher, J. Ehley, E. Grothbeck, J. Hammes, L. Hammes, J. Hobbes, D. Hornung, E. Horowitz, J. Johnston, H. Jones, M. Kamari, J. Katcha, A. Kelly, H. Kijenski, H. Kirkpatrick, N. Kramer, J. Marceille, R. Mohamad-Hassan, E. Mohd Nazir, Z. Othman, L. Pitcher, H. Rahman, P. Scherckenback, K. Streif, W. Vandewalle and R. Williams.

REFERENCES

- 1 Daniels, D.L., Plunkett, G.I., III, Burland, V. and Blattner, F.R., (1992) *Science*, **257**, 771–778.
- 2 Burland, V., Plunkett, G.I., III, Daniels, D.L. and Blattner, F.R., (1993) *Genomics*, **16**, 551–561.
- 3 Plunkett, G.I., III, Burland, V., Daniels, D.L. and Blattner, F.R., (1993) *Nucleic Acids Res.*, **21**, 3391–3398.
- 4 Blattner, F.R., Burland, V., Plunkett, G.I., III, Sofia, H.J. and Daniels, D.L., (1993) *Nucleic Acids Res.*, **21**, 5408–5417.
- 5 Sofia, H.J., Burland, V., Daniels, D.L., Plunkett, G., III and Blattner, F.R., (1994) *Nucleic Acids Res.*, **22**, 2576–2586.
- 6 Yura, T., Mori, H., Nagai, H., Nagata, T., Ishihama, A., *et al.*, (1992) *Nucleic Acids Res.*, **20**, 3305–3308.

- 7 Fujita, N., Mori, H., Yura, T. and Ishihama, A., (1994) *Nucleic Acids Res.*, **22**, 1637–1639.
- 8 Daniels, D.L. (1990) *Constructing Encyclopedias of Genomes*, in *The Bacterial Chromosome*, K. Drlica and M. Riley (eds) American Society for Microbiology, Washington, DC. p. 43–51.
- 9 Burland, V., Daniels, D.L., Plunkett, G.I.III, and Blattner, F.R., (1993) *Nucleic Acids Res.*, **21**, 3385–3390.
- 10 Bachmann, B.J., (1990) *Microbiology Rev.*, **54**, 130–197.
- 11 Bachi, B. and Kornberg, H.L., (1975) *J. Gen. Microbiol.*, **90**, 321–335.
- 12 Isturiz, T., Palmero, E. and Vitelli-Flores, J., (1986) *J. Gen. Microbiol.*, **132**, 3209–3219.
- 13 Fujita, Y., Fujita, T., Miwa, Y., Nihashi, J. and Aratani, Y., (1986) *J. Biol. Chem.*, **261**, 13744–13753.
- 14 Zhu, N. and Roth, J.R., (1990) *J. Bacteriol.*, **173**, 1302–1310.
- 15 Foster, J.W. and Penfound, T., (1993) *FEMS Microbiol. Lett.*, **112**, 179–184.
- 16 Foster, J.W., Park, Y.K., Penfound, T., Fenger, T. and Spector, M.P., (1990) *J. Bacteriol.*, **172**, 4187–4196.
- 17 Tobe, T., Sasakawa, C., Okada, N., Honma, Y. and Yoshikawa, M., (1992) *J. Bacteriol.*, **174**, 6359–6367.
- 18 Wolfe, S.A. and Smith, J.M., (1988) *J. Biol. Chem.*, **263**, 19147–19153.
- 19 Robert-Baudouy, J., Portalier, R. and Stoeber, F., (1981) *J. Bacteriol.*, **145**, 211–220.
- 20 Ritzenthaler, P. and Mata-Gilsinger, M., (1982) *J. Bacteriol.*, **150**, 1040–1047.
- 21 Blanco, C., Ritzenthaler, P. and Kolb, A., (1986) *Mol. Gen. Genet.*, **202**, 112–119.
- 22 Ritzenthaler, P., Mata-Gilsinger, M. and Stoeber, F., (1980) *J. Bacteriol.*, **143**, 1116–1126.
- 23 Robert-Baudouy, J.M. and Portalier, R.C., (1974) *Mol. Gen. Genet.*, **131**, 31–46.
- 24 Blum, P.H., Jovanovich, S.B., McCann, M.P., Schultz, J.E., Lesley, S.A., *et al.*, (1990) *J. Bacteriol.*, **172**, 3813–3820.
- 25 Schultz, J.E. and Matin, A., (1991) *J. Mol. Biol.*, **218**, 129–140.
- 26 Clyman, J., Quirk, S. and Belfort, M. (1994) *Mobile Introns in the T-Even Phages*, in *Molecular Biology of Bacteriophage T4*, J.D. Karam (ed.) American Society for Microbiology, Washington, D.C. p. 83–88.
- 27 von Heijne, G., (1986) *Nucleic Acids Res.*, **14**, 4683–4690.
- 28 Link, A.J., (1994) thesis. Harvard University.
- 29 Burge, C., Campbell, A.M. and Karlin, S., (1992) *Proc. Natl. Acad. Sci. USA*, **89**, 1358–1362.
- 30 Karlin, S. and Cardon, L.R., (1994) *Ann. Rev. Microbiol.*, **48**, 619–654.
- 31 McClelland, M., (1985) *J. Mol. Evolution*, **21**, 317–322.
- 32 McClelland, M., Jones, R., Patel, Y. and Nelson, M., (1987) *Nucleic Acids Res.*, **15**, 5985–6005.
- 33 Médigue, C., Rouxel, T., Vigier, P., Hénaut, A. and Danchin, A., (1991) *J. Mol. Biol.*, **222**, 851–856.
- 34 McClelland, M. and Bhagwat, A.S. (1992) *Nature*, **355**, 595–596.
- 35 Bhagwat, A.S. and McClelland, M. (1992) *Nucleic Acids Res.*, **20**, 1663–1668.
- 36 Merkl, R., Kroger, M., Rice, P. and Fritz, H.J. (1992) *Nucleic Acids Res.*, **20**, 1657–1662.
- 37 Chen, G.-F.T. and Inouye, M., (1994) *Genes Dev.*, **8**, 2641–2652.
- 38 Wanner, B.L., (1993) *J. Cell. Biochem.*, **51**, 47–54.
- 39 Makino, K., Kim, S., Shinagawa, H., Amemura, M. and Nakata, A., (1991) *J. Bacteriol.*, **173**, 2665–2672.
- 40 Metcalf, W.W. and Wanner, B.L., (1993) *J. Bacteriol.*, **175**, 3430–3442.
- 41 Metcalf, W.W. and Wanner, B.L., (1993) *Gene*, **129**, 27–32.
- 42 Murray, N.E., Daniel, A.S., Cowan, G.M. and Sharp, P.M., (1993) *Mol. Microbiol.*, **9**, 133–143.
- 43 Greener, T., Govezensky, D. and Zamir, A., (1993) *EMBO J.*, **12**, 889–896.
- 44 Lanfroy, E. and Bohin, J.-P. (1993) *J. Bacteriol.*, **175**, 5736–5737.
- 45 Berlyn, M.B. and Letovsky, S., (1992) *Nucleic Acids Res.*, **20**, 6143–6151.
- 46 Fuchs, R., (1991) *Comput. Applic. Biosci.*, **7**, 105–106.
- 47 Staudenmaier, H., Van Hove, B., Yaraghi, Z. and Braun, V., (1989) *J. Bacteriol.*, **171**, 2626–2633.
- 48 Klemm, P., Krogfelt, K.A., Hedegaard, L. and Christiansen, G., (1990) *Mol. Microbiol.*, **4**, 553–559.
- 49 Jones, C.H., Pinkner, J.S., Nicholes, A.V., Slonim, L.N., Abraham, S.N., *et al.*, (1993) *Proc. Natl. Acad. Sci. USA*, **90**, 8397–8401.
- 50 Rossolini, G.M., Muscas, P., Chiesurin, A. and Satta, G., (1993) *FEMS*, **114**, 259–266.
- 51 Klemm, P. and Christiansen, G., (1990) *Mol. Gen. Genet.*, **220**, 334–338.
- 52 Nikaido, H. and Normark, S., (1987) *Mol. Microbiol.*, **1**, 29–36.
- 53 Amemura, M., Makino, K., Shinagawa, H. and Nakata, A., (1986) *J. Bacteriol.*, **168**, 294–302.
- 54 Yim, H.H. and Villarejo, M., (1992) *J. Bacteriol.*, **174**, 3637–3644.
- 55 Crooke, H.R. (1994) GenBank accession no. X77707.
- 56 Liu, J., Burns, D.M. and Beacham, I.R., (1986) *J. Bacteriol.*, **165**, 1002–1010.
- 57 Neuwald, A.F. and Stauffer, G.V., (1989) *Gene*, **82**, 219–228.
- 58 Tsui, H.-C.T., Zhao, G., Feng, G., Leung, H.-C.E. and Winkler, M.E., (1994) *Mol. Microbiol.*, **11**, 189–202.
- 59 Engel, H., Kazemier, B. and Keck, W., (1991) *J. Bacteriol.*, **173**, 6773–6782.
- 60 Westenber, D.J., Gunsalus, R.P., Ackrell, B.A.C. and Cecchini, G., (1990) *J. Biol. Chem.*, **265**, 19560–19567.