Biochem. J. (1984) **224**, 799–815
Printed in Great Britain

DNA sequence around the *Escherichia coli unc* operon

Completion of the sequence of a 17 kilobase segment containing *asn*A, *ori*C, *unc*, *glm*S and *pho*S

John E. WALKER,* Nicholas J. GAY, Matti SARASTE† and Alex N. EBERLE‡
Medical Research Council Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, U.K.

(*Received 13 August 1984/Accepted 29 August 1984*)

The nucleotide sequence is described of a region of the *Escherichia coli* chromosome extending from *ori*C to *pho*S that also includes the loci *gid*, *unc* and *glm*S. Taken with known sequences for *asn*A and *pho*S this completes the sequence of a segment of about 17 kilobases or 0.4 min of the *E. coli* genome. Sequences that are probably transcriptional promoters for *unc* and *pho*S can be detected and the identity of the *unc* promoter has been confirmed by experiments *in vitro* with RNA polymerase. Upstream of the promoter sequence is an extensive region that appears to be non-coding. Conserved sequences are found that may serve to concentrate RNA polymerase in the vicinity of the *unc* promoter. Hairpin loop structures resembling known rho-independent transcription termination signals are evident following the *unc* operon and *glm*S. The *glm*S gene encoding the amidotransferase, glucosamine synthetase, has been identified by homology with glutamine 5-phosphoribosylpyrophosphate amidotransferase.

The *Escherichia coli unc* operon, encoding the eight subunits of ATP synthase, is found near min 83 in the 100 min linkage map close to the single origin of bidirectional DNA replication, *ori*C (Bachman, 1983). The region between *ori*C and *unc* is potentially of interest because of its proximity to the origin of replication. Earlier it had been suggested that an outer membrane protein binding at or near the origin of replication might be encoded in this region of the chromosome, or alternatively that the DNA segment to which the outer membrane protein is thought to bind might lie between *ori*C and *unc* (Wolf-Watz & Norquist, 1979; Wolf-Watz & Masters, 1979). The phenotypic marker *het* has been used for this trait (see Bachman, 1983). Recently two DNA segments have been proposed to bind to the membrane protein; one overlaps *ori*C, the other lies within the *unc* operon (Wolf-Watz, 1984). A second phenotypic trait *gid* (glucose-inhibited division) has also been associated with the region between *ori*C and *unc*. This phenotype was designated following the construction of strains carrying a deletion of *ori*C and part of *gid* and with an insertion of transposon Tn10 in *asn*A. This *ori*C deletion strain can be maintained by replication of an integrated F-plasmid. Various *ori*C minichromosomes were integrated into the *ori*C deletion strain by homologous recombination. It was observed that integrated minichromosomes carrying an intact *gid*A gene displayed a 30% higher growth rate on glucose media compared with ones in which *gid*A was partly or completely absent. A protein of 70 kDa has been associated with *gid*A (von Meyenberg & Hansen, 1980). Insertion of transposon Tn10 in *gid*A also influences expression of a 25 kDa protein, the gene for which maps between *gid*A and *unc*. Therefore, it has been proposed that the 70 kDa and 25 kDa proteins are co-transcribed; *gid*B has been used to designate the gene for the 25 kDa protein (von Meyenberg *et al.*, 1982).

The DNA sequence of this region of the *E. coli* genome is described here. It confirms the presence of two genes that would encode proteins of the sizes determined for the *gid*A and *gid*B gene products. They possibly comprise a single transcriptional unit. These genes are followed by an extensive region of DNA that appears to be non-coding. It contains within it the transcriptional promoter of

Abbreviations used: kb, kilobases; bp, base pairs.

* To whom correspondence and requests for reprints should be addressed.

† Present address: Department of Medical Chemistry, University of Helsinki, Finland.

‡ Present address: Research Department (ZLF), Kantonsspital, Hebelstrasse 20, CH-4031 Basel, Switzerland.

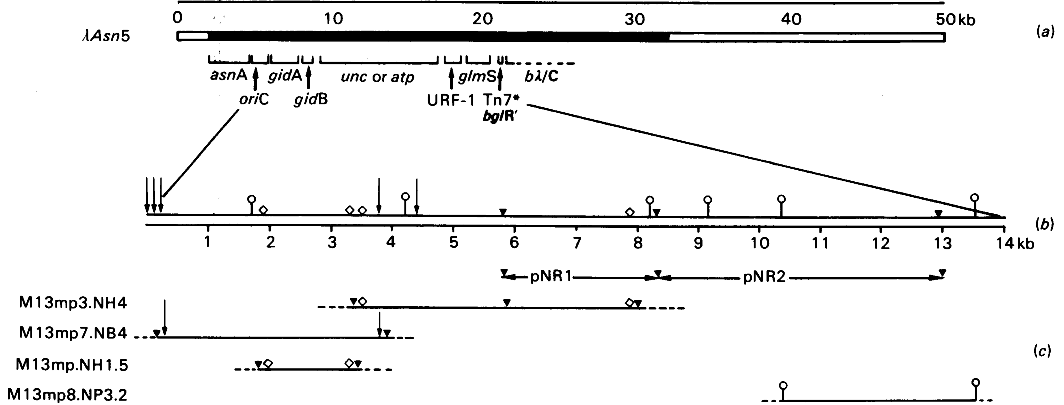


Fig. 1. *Genetic and physical maps in the region of the unc operon*
(*a*) Extent of *E. coli* chromosome in transducing phage *λAsn*5 (black) with genetic markers. (*b*) Restriction sites (redrawn from Futai & Kanazawa, 1980): ↓, *Bam*H1; ♀, *Pst*I; ◇, *Hin*dIII; ▼, *Eco*RI. (*c*) Location of sub-cloned fragments.

the *unc* operon. We have described the DNA sequence of the *unc* operon previously and parts of it have been independently determined. The operon contains nine genes. (For a review see Walker *et al.*, 1984.)

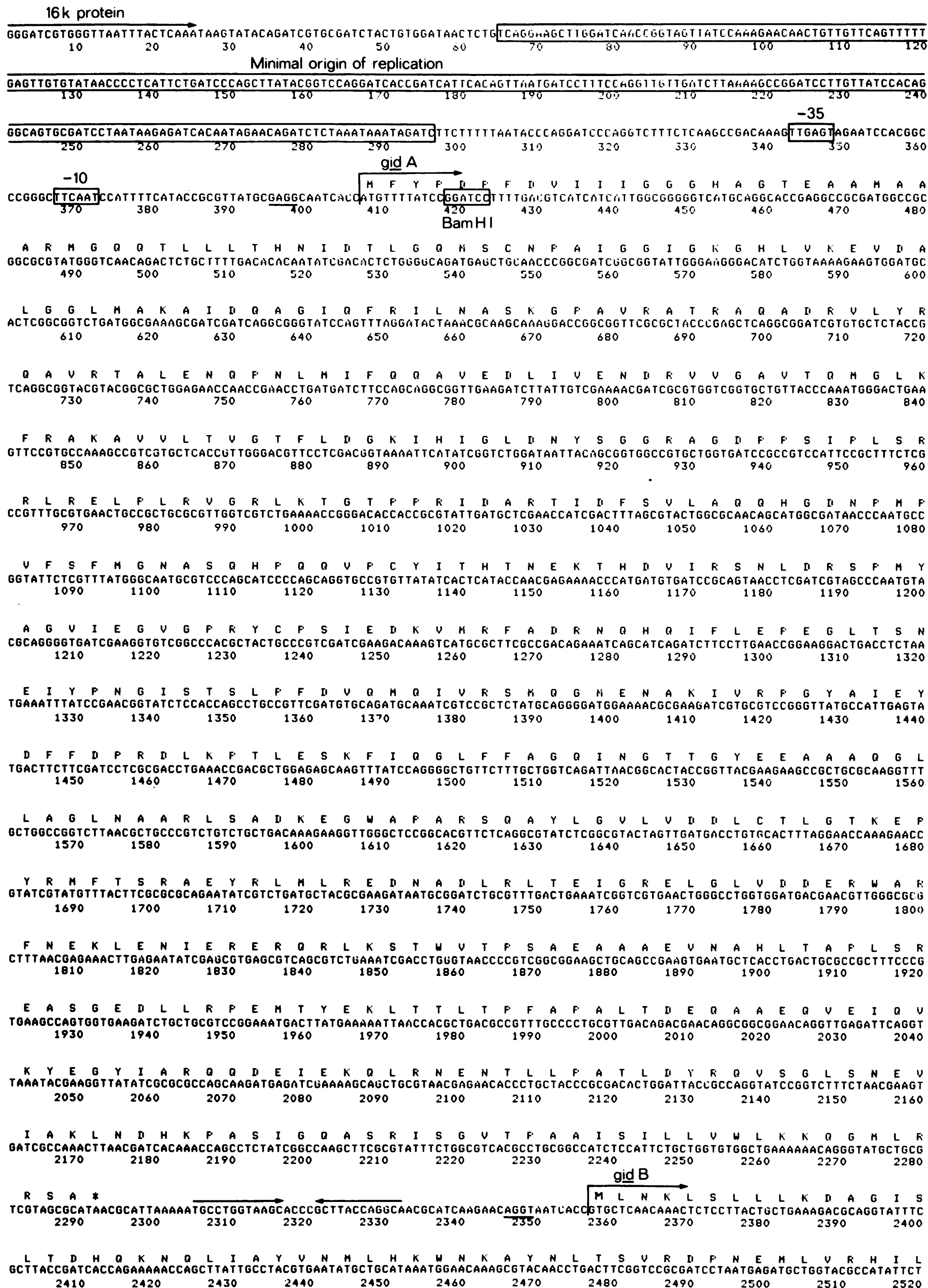
We also describe here the DNA sequence of about 2.2 kb beyond the *unc* operon that joins the sequence of the *unc* operon to that of *pho*S. It contains two genes which may be co-transcribed. The genetic marker *glm*S corresponding to the amidotransferase glucosamine-6-phosphate synthetase maps in this region (Wu & Wu, 1971; Bachman, 1983). The protein sequence predicted from the DNA sequence of the second of these two reading frames has striking homology with another amidotransferase, 5-phosphoribosylamine: glutamine pyrophosphate phosphoribosyltransferase, the product of *pur*F in both *E. coli* (Tso *et al.*, 1982*a*) and *Bacillus subtilis* (Makaroff *et al.*, 1983). On this basis *glm*S has been assigned to this second reading frame. The first of these two open reading frames (called here *Eco*URF-1) remains unidentified. As discussed elsewhere (N. J. Gay, V. L. J. Tybulewicz & J. E. Walker, unpublished work), the DNA sequence in this region also shows that the transposable element Tn7 inserts into a structure in the DNA sequence that appears to be the rho-independent transcriptional terminator for *glm*S. Taken with published DNA sequences for *asn*A (Nakamura *et al.*, 1981), *ori*C (Meijer *et al.*, 1979; Sugimoto *et al.*, 1979), the Tn7 insertion site (Lichtenstein & Brenner, 1982) and *pho*S (Magota *et al.*, 1984; Surin *et al.*, 1984), the sequences presented here complete the sequence of a segment of approx. 17 kb or 0.4 min of the *E. coli* genome.

Materials and methods

Cloning and preparation of DNA

The lysogen bacteriophage *λAsn*5 contains about 26 kb of the *E. coli* chromosome including the genetic loci *asn*, *ori*C, *unc* and *glm*S (Kanazawa *et al.*, 1980). Phage was grown up and DNA extracted from it as described previously (Gay & Walker, 1981*a*). The region from *ori*C to the beginning of the *unc* operon was cloned into bacteriophage M13. Three such recombinants that together cover this region (see Fig. 1), namely M13mp3.NH4, M13mp7.NB4 and M13mp3.NH1.5, were isolated and replicative form was prepared as described previously (Gay & Walker, 1981*a*,*b*). A region extending from the beginning of *unc*D to the *glm*S region is contained in a 4.5 kb *Eco*RI fragment (called R2) that had earlier been subcloned into the plasmid pACYC184 (Gay & Walker, 1981*a*). This recombinant plasmid is pNR2. Its preparation has been described in earlier work (Gay & Walker, 1981*a*). A *Pst*I–*Eco*RI fragment (extending from nucleotide 10679 to nucleotide 13556) was prepared as described by Saraste *et al.* (1981). It was then self-ligated, sonicated and, after end repair (Deininger, 1983), blunt-end-ligated into M13mp8 (Messing & Vieira, 1982) under standard conditions, and finally used to transfect *E. coli* JM101.

Details of other methods concerning transformation and transfection have been given before (Walker & Gay, 1983). Single-stranded M13 templates for DNA sequencing were prepared under standard conditions. DNA sequences were determined by the dideoxy chain termination method (Sanger *et al.*, 1977) using a 17-nucleotide-


16k protein
Minimal origin of replication
-35
-10
gid A
BamH I
gid B

D S I V V A P Y L Q G E R F I D V G T G P G L P G I F L S I V R P E A H F T L L

CGATAGCATTGTGGTGGCACCGTATCTGCAAGGTGAACGGTTTATCGATGTCGGCACCGGACCAGGACTGCCAGGCATTCCACTCTCTATCGTGCGTCCTGAAGCCCATTTCACTCTGTT

2530 2540 2550 2560 2570 2580 2590 2600 2610 2620 2630 2640

D S L G K R V R F L R Q V Q H E L K L E N I E P V Q S R V E E F P S E P P F D G

GGATAGCCTTGGTAAACGCGTGCGTTTCCTTCGTCAGGTGCAACATGAGCTTAAACTGGAGAATATTGAACCAGTACAGAGCAGGGTAGAAGAGTTTCCTTCAGAGCCGCCATTTGATGG

2650 2660 2670 2680 2690 2700 2710 2720 2730 2740 2750 2760

V I S R A F A S L N D M V S W C H H L P G E Q G R F Y A L K G Q M P E D E I A L

CGTAATTAGCCGCGCTTTTGCCTCTCTGAACGATATGGTGAGCTGGTGCCACCATCTTCCTGGTGAGCAAGGCCGTTTCTACGCGCTGAAAGGGCAAATGCCGGAAGATGAAATCGCTTT

2770 2780 2790 2800 2810 2820 2830 2840 2850 2860 2870 2880

L P E E Y Q V E S V V K L Q V P A L D G E R H L V V I K A N K I *

GTTGCCCGAAGAATATCAGGTCGAATCAGTGGTTAAACTTCAGGTTCCAGCCCTGGATGGCGAACGTCATCTGGTGGTGATTAAAGCAAATAAAATTTAATTTTTATCAAAAAAATCATA

2890 2900 2910 2920 2930 2940 2950 2960 2970 2980 2990 3000

AAAAATTGACCGGTTAGACTGTTAACAACAACCAGGTTTTCTACTGATATAACTGGTTACATTTAACGCCACGTTCACTCTTTTGCATCAACAAGATAACGTGGCTTTTTTTGGTAAGCA

3010 3020 3030 3040 3050 3060 3070 3080 3090 3100 3110 3120

GAAAATAAGTCATTAGTGAAAATATCAGTCTGCTAAAAATCGGCGCTAAGAACCATCATTGGCTGTTAAAACATTATTAAAAATGTCAATGGGTGGTTTTTGTTGTGTAAATGTCATTTA

3130 3140 3150 3160 3170 3180 3190 3200 3210 3220 3230 3240

locator

TTAAAACAGTATCTGTTTTTAGACTGAAATATCATAAACTTGCAAAGGCATCATTTGCCAAGTAAATAAATATGCTGTGCGCGAACATGCGCAATATGTGATCTGAAGCACGCTTTATCA

3250 3260 3270 3280 3290 3300 3310 3320 3330 3340 3350 3360

locator locator locator locator p-35

CCAGTGTTTACGCGTTATTTACAGTTTTTCATGATCGAACAGGGTTAGCAGAAAAGTCGCAATTGTATGCACTGGAAAAATATTTAAACATTTATTCACCTTTTGGCTACTTATTGTTTG

3370 3380 3390 3400 3410 3420 3430 3440 3450 3460 3470 3480

-10 1 *unc* I

M K N V M S

AAATCACGGGGGCGCACCGTATAATTTGACCGCTTTTTGATGCTTGACTCTAAGCCTTAAAGAAAGTTTTATACGACACGCGGCATACCTCGAAGGGAGCAGGAGTGAAAAACGTGATGT

3490 3500 3510 3520 3530 3540 3550 3560 3570 3580 3590 3600

V S L V S R N V A R K L L L V Q L L V V I A S G L L F S L K D P F W G V S A I S

CTGTGTCGCTCGTGAGTCGAAACGTTGCTCGGAAGCTTCTGCTCGTTCAGTTACTGGTGGTGATAGCAAGTGGATTGCTGTTCAGCCTCAAAGACCCCTTCTGGGGCGTCTCTGCAATAA

3610 3620 3630 3640 3650 3660 3670 3680 3690 3700 3710 3720

G G L A V F L P N V L F M I F A W R H Q A H T P A K G R V A W T F A F G E A F K

GCGGGGGCCTGGCAGTCTTTCTGCCTAACGTTTTGTTTATGATATTTGCCTGGCGTCACCAGGCGCATACACCAGCGAAAGGCCGGGTGGCCTGGACATTCGCATTTGGCGAAGCTTTCA

3730 3740 3750 3760 3770 3780 3790 3800 3810 3820 3830 3840

V L A M L V L L V V A L A V L K A V F L P L I V T W V L V L V V Q I L A P A V I

AAGTTCTGGCGATGTTGGTGTTACTGGTGGTGGCGTTGGCGGTTTTAAAGGCGGTATTCTTGCCGCTGATCGTTACGTGGGTTTTGGTGCTGGTGGTTCAGATACTGGCACCGGCTGTAA

3850 3860 3870 3880 3890 3900 3910 3920 3930 3940 3950 3960

unc B (a)

M A S E N M T P Q D Y I G H H L N N L Q L D L R T F S L V D P Q

N N K G *

TTAACAACAAAGGGTAAAAGGCATCATGGCTTCAGAAAATATGACGCCGCAGGATTACATAGGACACCACCTGAATAACCTTCAGCTGGACCTGCGTACATTCTCGCTGGTGGATCCACA

3970 3980 3990 4000 4010 4020 4030 4040 4050 4060 4070 4080

N P P A T F W T I N I D S M F F S V V L G L L F L V L F R S V A K K A T S G V P

AAACCCCCCAGCCACCTTCTGGACAATCAATATTGACTCCATGTTCTTCTCGGTGGTGCTGGGTCTGTTGTTCCTGGTTTTATTCCGTAGCGTAGCCAAAAAGGCGACCAGCGGTGTGCC

4090 4100 4110 4120 4130 4140 4150 4160 4170 4180 4190 4200

G K F Q T A I E L V I G F V N G S V K D M Y H G K S K L I A P L A L T I F V W V

AGGTAAGTTTCAGACCGCGATTGAGCTGGTGATCGGCTTTGTTAATGGTAGCGTGAAAGACATGTACCATGGCAAAAGCAAGCTGATTGCTCCGCTGGCCCTGACGATCTTCGTCTGGGT

4210 4220 4230 4240 4250 4260 4270 4280 4290 4300 4310 4320

F L M N L M D L L P I D L L P Y I A E H V L G L P A L R V V P S A D V N V T L S

ATTCCTGATGAACCTGATGGATTTACTGCCTATCGACCTGCTGCCGTACATTGCTGAACATGTACTGGGTCTGCCTGCACTGCGTGTGGTTCCGTCTGCGGACGTGAACGTAACGCTGTC

4330 4340 4350 4360 4370 4380 4390 4400 4410 4420 4430 4440

M A L G V F I L I L F Y S I K M K G I G G F T K E L T L Q P F N H W A F I P V N

TATGGCACTGGGCGTATTTATCCTGATTCTGTTCTACAGCATCAAAATGAAAGGCATCGGCGGCTTCACGAAAGAGTTGACGCTGCAGCCGTTCAATCACTGGGCGTTCATTCCTGTCAA

4450 4460 4470 4480 4490 4500 4510 4520 4530 4540 4550 4560

L I L E G V S L L S K P V S L G L R L F G N M Y A G E L I F I L I A G L L P W W

CTTAATCCTTGAAGGGGTAAGCCTGCTGTCCAAACCAGTTTCACTCGGTTTGCGACTGTTCGGTAACATGTATGCCGGTGAGCTGATTTTCATTCTGATTGCTGGTCTGTTGCCGTGGTG

4570 4580 4590 4600 4610 4620 4630 4640 4650 4660 4670 4680

S Q W I L N V P W A I F H I L I I T L Q A F I F M V L T I V Y L S M A S E E H *

GTCACAGTGGATCCTGAATGTGCCGTGGGCCATTTTCCACATCCTGATCATTACGCTGCAAGCCTTCATCTTCATGGTTCTGACGATCGTCTATCTGTCGATGGCGTCTGAAGAACATTA

4690 4700 4710 4720 4730 4740 4750 4760 4770 4780 4790 4800

unc E (c)

M E N L N M D L L Y M A A A V M M G L A A I G A A

ATTTACCAACACTACTACGTTTTAACTGAAACAAACTGGAGACTGTCATGGAAAACCTGAATATGGATCTGCTGTACATGGCTGCCGCTGTGATGATGGGTCTGGCGGCAATCGGTGCTG

4810 4820 4830 4840 4850 4860 4870 4880 4890 4900 4910 4920

I G I G I L G G K F L E G A A R Q P D L I P L L R T Q F F I V M G L V D A I P M
CGATCGGTATCGGCATCCTCGGGGGTAAATTCCTGGAAGGCGCAGCGCGTCAACCTGATCTGATTCCTCTGCTGCGTACTCAGTTCTTTATCGTTATGGGTCTGGTGGATGCTATCCCGA
4930 4940 4950 4960 4970 4980 4990 5000 5010 5020 5030 5040

unc F (b)

I A V G L G L Y V M F A V A * * M N L N
TGATCGCTGTAGGTCTGGGTCTGTACGTGATGTTCGCTGTCGCGTAGTAAGCGTTGCTTTTATTTAAAGAGCAATATCAGAACGTTAACTAAATAGAGGCATTGTGCTGTGAATCTTAAC
5050 5060 5070 5080 5090 5100 5110 5120 5130 5140 5150 5160

A T I L G Q A I A F V L F V L F C M K Y V W P P L M A A I E K R Q K E I A D G L
GCAACAATCCTCGGCCAGGCCATCGCGTTTGTCCTGTTCGTTCTGTTCTGCATGAAGTACGTATGGCCGCCATTAATGGCAGCCATCGAAAAACGTCAAAAAGAAATTGCTGACGGCCTT
5170 5180 5190 5200 5210 5220 5230 5240 5250 5260 5270 5280

A S A E R A H K D L D L A K A S A T D Q L K K A K A E A Q V I I E Q A N K R R S
GCTTCCGCAGAACGAGCACATAAGGACCTTGACCTTGCAAAGGCCAGCGCGACCGACCAGCTGAAAAAAGCGAAAGCGGAAGCCCAGGTAATCATCGAGCAGGCGAACAAACGCCGCTCG
5290 5300 5310 5320 5330 5340 5350 5360 5370 5380 5390 5400

Q I L D E A K A E A E Q E R T K I V A Q A Q A E I E A E R K R A R E E L R K Q V
CAGATTCTGGACGAAGCGAAAGCTGAGGCAGAACAGGAACGTACTAAAATCGTGGCCCAGGCGCAGGCGGAAATTGAAGCCGAGCGTAAACGTGCCCGTGAAGAGCTGCGTAAGCAAGTT
5410 5420 5430 5440 5450 5460 5470 5480 5490 5500 5510 5520

unc H (δ)

A I L A V A G A E K I I E R S V D E A A N S D I V D K L V A E L * M S E
GCTATCCTGGCTGTTGCTGGCGCCGAGAAGATCATCGAACGTTCCGTGGATGAAGCTGCTAACAGCGACATCGTGGATAAACTTGTCGCTGAACTGTAAGGAGGGAGGGGCTGATGTCTG
5530 5540 5550 5560 5570 5580 5590 5600 5610 5620 5630 5640

F I T V A R P Y A K A A F D F A V E H Q S V E R W Q D M L A F A A E V T K N E Q
AATTTATTACGGTAGCTCGCCCCTACGCCAAAGCAGCTTTTGACTTTGCCGTCGAACACCAAAGTGTAGAACGCTGGCAGGACATGCTGGCGTTTGCCGCCGAGGTAACCAAAAACGAAC
5650 5660 5670 5680 5690 5700 5710 5720 5730 5740 5750 5760

M A E L L S G A L A P E T L A E S F I A V C G E Q L D E N G Q N L I R V M A E N
AAATGGCAGAGCTTCTCTCTGGCGCGCTTGCGCCAGAAACGCTCGCCGAGTCGTTTATCGCAGTTTGTGGTGAGCAACTGGACGAAAACGGTCAGAACCTGATTCGGGTTATGGCTGAAA
5770 5780 5790 5800 5810 5820 5830 5840 5850 5860 5870 5880

G R L N A L P D V L E Q F I H L R A V S E A T A E V D V I S A A A L S E Q Q L A
ATGGTCGTCTTAACGCGCTCCCGGATGTTCTGGAGCAGTTTATTCACCTGCGTGCCGTGAGTGAGGCTACCGCTGAGGTAGACGTCATTTCCGCTGCCGCACTGAGTGAACAACAGCTCG
5890 5900 5910 5920 5930 5940 5950 5960 5970 5980 5990 6000

K I S A A M E K R L S R K V K L N C K I D K S V M A G V I I R A G D M V I D G S
CGAAAATTTCTGCTGCGATGGAAAAACGTCTGTCACGCAAAGTTAAGCTGAATTGCAAAATCGATAAGTCTGTAATGGCAGGCGTTATCATCCGAGCGGGTGATATGGTCATTGATGGCA
6010 6020 6030 6040 6050 6060 6070 6080 6090 6100 6110 6120

unc A (α)

V R G R L E R L A D V L Q S * M Q L N S T E I S E L I K Q R I A Q F N V
GCGTACGCGGTCGTCTTGAGCGCCTTGCAGACGTCTTGCAGTCTTAAGGGGACTGGAGCATGCAACTGAATTCCACCGAAATCAGCGAACTGATCAAGCAGCGCATTGCTCAGTTCAATG
6130 6140 6150 6160 6170 6180 6190 6200 6210 6220 6230 6240

V S E A H N E G T I V S V S D G V I R I H G L A D C M Q G E M I S L P G N R Y A
TTGTGAGTGAAGCTCACAACGAAGGTACTATTGTTTCTGTAAGTGACGGTGTTATCCGCATTCACGGCCTGGCCGATTGTATGCAGGGTGAAATGATCTCCCTGCCGGGTAACCGTTACG
6250 6260 6270 6280 6290 6300 6310 6320 6330 6340 6350 6360

I A L N L E R D S V G A V V M G P Y A D L A E G M K V K C T G R I L E V P V G R
CTATCGCACTGAACCTCGAGCGCGACTCTGTAGGTGCGGTTGTTATGGGTCCGTACGCTGACCTTGCCGAAGGCATGAAAGTTAAGTGTACTGGCCGTATCCTGGAAGTTCCGGTTGGCC
6370 6380 6390 6400 6410 6420 6430 6440 6450 6460 6470 6480

G L L G R V V N T L G A P I D G K G P L D H D G F S A V E A I A P G V I E R Q S
GTGGCCTGCTGGGCCGTGTGGTTAACACTCTGGGTGCACCAATCGACGGTAAAGGTCCGCTGGATCACGACGGCTTCTCTGCTGTAGAAGCAATCGCTCCGGGCGTTATCGAACGTCAGT
6490 6500 6510 6520 6530 6540 6550 6560 6570 6580 6590 6600

V D Q P V Q T G Y K A V D S M I P I G R G Q R E L I I G D R Q T G K T A L A I D
CCGTAGATCAGCCGGTACAGACCGGTTATAAAGCCGTTGACTCCATGATCCCAATCGGTCGTGGTCAGCGTGAATTGATCATCGGTGACCGTCAGACAGGTAAAACCGCACTGGCTATCG
6610 6620 6630 6640 6650 6660 6670 6680 6690 6700 6710 6720

A I I N Q R D S G I K C I Y V A I G Q K A S T I S N V V R K L E E H G A L A N T
ATGCCATCATCAACCAGCGCGATTCCGGTATCAAATGTATCTATGTCGCTATCGGCCAGAAAGCGTCCACCATTTCTAACGTGGTACGTAAACTGGAAGAGCACGGCGCACTGGCTAACA
6730 6740 6750 6760 6770 6780 6790 6800 6810 6820 6830 6840

I V V V A T A S E S A A L Q Y L A R M P V A L M G E Y F R D R G E D A L I I Y D
CCATCGTTGTGGTAGCAACCGCGTCTGAATCCGCTGCACTGCAATACCTGGCACGTATGCCGGTTGCGCTAATGGGCGAATACTTCCGTGACCGCGGTGAAGATGCGCTGATCATTTACG
6850 6860 6870 6880 6890 6900 6910 6920 6930 6940 6950 6960

D L S K Q A V A Y R Q I S L L L R R P P G R E A F P G D V F Y L H S R L L E R A
ATGACCTGTCTAAACAGGCTGTTGCTTACCGTCAGATCTCCCTGCTGCTCCGTCGTCCGCCAGGACGTGAAGCATTCCCGGGCGACGTTTTCTACCTCCACTCTCGTCTGCTGGAGCGTG
6970 6980 6990 7000 7010 7020 7030 7040 7050 7060 7070 7080

A R V N A E Y V E A F T K G E V K G K T G S L T A L P I I E T Q A G D V S A F V
CTGCACGTGTTAACGCCGAATACGTTGAAGCCTTCACCAAAGGTGAAGTGAAAGGGAAAACCGGTTCTCTGACCGCACTGCCGATTATCGAAACTCAGGCGGGTGACGTTTCTGCGTTCG
7090 7100 7110 7120 7130 7140 7150 7160 7170 7180 7190 7200

P T N V I S I T D G Q I F L E T N L F N A G I R P A V N P G I S V S R V G G A A
TTCCGACCAACGTAATCTCCATTACCGATGGTCAGATCTTCCTGGAAACCAACCTGTTCAACGCCGGTATTCGTCCTGCGGTTAACCCGGGTATTTCCGTATCCCGTGTTGGTGGTGCAG
7210 7220 7230 7240 7250 7260 7270 7280 7290 7300 7310 7320

```
  Q  T  K  I  M  K  K  L  S  G  G  I  R  T  A  L  A  Q  Y  R  E  L  A  A  F  S  Q  F  A  S  D  L  D  D  A  T  R  K  Q  L
CACAGACCAAGATCATGAAAAAACTGTCCGGTGGTATCCGTACCGCTCTGGCACAGTATCGTGAACTGGCAGCGTTCTCTCAGTTTGCATCCGACCTTGACGATGCAACACGTAAGCAGC
      7330      7340      7350      7360      7370      7380      7390      7400      7410      7420      7430      7440

  D  H  G  Q  K  V  T  E  L  L  K  Q  K  Q  Y  A  P  M  S  V  A  Q  Q  S  L  V  L  F  A  A  E  R  G  Y  L  A  D  V  E  L
TTGACCACGGTCAGAAAGTGACCGAACTGCTGAAACAGAAACAGTATGCGCCGATGTCCGTTGCGCAGCAGTCTCTGGTTCTGTTCGCAGCAGAACGTGGTTACCTGGCGGATGTTGAAC
      7450      7460      7470      7480      7490      7500      7510      7520      7530      7540      7550      7560

  S  K  I  G  S  F  E  A  A  L  L  A  Y  V  D  R  D  H  A  P  L  M  Q  E  I  N  Q  T  G  G  Y  N  D  E  I  E  G  K  L  K
TGTCGAAAATTGGCAGCTTCGAAGCCGCTCTGCTGGCTTACGTCGACCGTGATCACGCTCCGTTGATGCAAGAGATCAACCAGACCGGTGGCTACAACGACGAAATCGAAGGCAAGCTGA
      7570      7580      7590      7600      7610      7620      7630      7640      7650      7660      7670      7680
                                                                                                 unc G (γ)
                                                                                                 ──────────►
                                                                                                  M  A  G  A  K  E  I  R  S  K
  G  I  L  D  S  F  K  A  T  Q  S  W  *
AAGGCATCCTCGATTCCTTCAAAGCAACCCAATCCTGGTAACGTCTGGCGGCTTGCCTTAGGGCAGGCCGCAAGGCATTGAGGAGAAGCTCATGGCCGGCGCAAAAGAGATACGTAGTAA
      7690      7700      7710      7720      7730      7740      7750      7760      7770      7780      7790      7800

  I  A  S  V  Q  N  T  Q  K  I  T  K  A  M  E  M  V  A  A  S  K  M  R  K  S  Q  D  R  M  A  A  S  R  P  Y  A  E  T  M  R
GATCGCAAGCGTCCAGAACACGCAAAAGATCACTAAAGCGATGGAGATGGTCGCCGCTTCCAAAATGCGTAAATCGCAGGATCGCATGGCGGCCAGCCGTCCTTATGCAGAAACCATGCG
      7810      7820      7830      7840      7850      7860      7870      7880      7890      7900      7910      7920

  K  V  I  G  H  L  A  H  G  N  L  E  Y  K  H  P  Y  L  E  D  R  D  V  K  R  V  G  Y  L  V  V  S  T  D  R  G  L  C  G  G
CAAAGTGATTGGTCACCTTGCACACGGTAATCTGGAATATAAGCACCCTTACCTGGAAGACCGCGACGTTAAACGCGTGGGCTACCTGGTGGTGTCGACCGACCGTGGTTTGTGCGGTGG
      7930      7940      7950      7960      7970      7980      7990      8000      8010      8020      8030      8040

  L  N  I  N  L  F  K  K  L  L  A  E  M  K  T  W  T  D  K  G  V  Q  C  D  L  A  M  I  G  S  K  G  V  S  F  F  N  S  V  G
TTTGAACATTAACCTGTTCAAAAAACTGCTGGCGGAAATGAAGACCTGGACCGACAAAGGCGTTCAATGCGACCTCGCAATGATCGGCTCGAAAGGCGTGTCGTTCTTCAACTCCGTGGG
      8050      8060      8070      8080      8090      8100      8110      8120      8130      8140      8150      8160

  G  N  V  V  A  Q  V  T  G  M  G  D  N  P  S  L  S  E  L  I  G  P  V  K  V  M  L  Q  A  Y  D  E  G  R  L  D  K  L  Y  I
CGGCAATGTTGTTGCCCAGGTCACCGGCATGGGGGATAACCCTTCCCTGTCCGAACTGATCGGTCCGGTAAAAGTGATGTTGCAGGCCTACGACGAAGGCCGTCTGGACAAGCTTTACAT
      8170      8180      8190      8200      8210      8220      8230      8240      8250      8260      8270      8280

  V  S  N  K  F  I  N  T  M  S  Q  V  P  T  I  S  Q  L  L  P  L  P  A  S  D  D  D  D  L  K  H  K  S  W  D  Y  L  Y  E  P
TGTCAGCAACAAATTTATTAACACCATGTCTCAGGTTCCGACCATCAGCCAGCTGCTGCCGTTACCGGCATCAGATGATGATGATCTGAAACATAAATCCTGGGATTACCTGTACGAACC
      8290      8300      8310      8320      8330      8340      8350      8360      8370      8380      8390      8400

  D  P  K  A  L  L  D  T  L  L  R  R  Y  V  E  S  Q  V  Y  Q  G  V  V  E  N  L  A  S  E  Q  A  A  R  M  V  A  M  K  A  A
CGATCCGAAGGCGTTGCTGGATACCCTGCTGCGTCGTTATGTCGAATCTCAGGTTTATCAGGGCGTGGTTGAAAACCTGGCCAGCGAGCAGGCCGCCCGTATGGTGGCGATGAAAGCCGC
      8410      8420      8430      8440      8450      8460      8470      8480      8490      8500      8510      8520

  T  D  N  G  G  S  L  I  K  E  L  Q  L  V  Y  N  K  A  R  Q  A  S  I  T  Q  E  L  T  E  I  V  S  G  A  A  A  V  *
GACCGACAATGGCGGCAGCCTGATTAAAGAGCTGCAGTTGGTATACAACAAAGCTCGTCAGGCCAGCATTACTCAGGAACTCACCGAGATCGTCTCGGGGGCCGCCGCGGTTTAAACAGG
      8530      8540      8550      8560      8570      8580      8590      8600      8610      8620      8630      8640
                          unc D (β)
                         ──────────►
                          M  A  T  G  K  I  V  Q  V  I  G  A  V  V  D  V  E  F  P  Q  D  A  V  P  R  V  Y  D  A  L  E  V  Q
TTATTTCGTAGAGGATTTAAGATGGCTACTGGAAAGATTGTCCAGGTAATCGGCGCCGTAGTTGACGTCGAATTCCCTCAGGATGCCGTACCGCGCGTGTACGATGCTCTTGAGGTGCAA
      8650      8660      8670      8680      8690      8700      8710      8720      8730      8740      8750      8760

  N  G  N  E  R  L  V  L  E  V  Q  Q  Q  L  G  G  G  I  V  R  T  I  A  M  G  S  S  D  G  L  R  R  G  L  D  V  K  D  L  E
AATGGTAATGAGCGTCTGGTGCTGGAAGTTCAGCAGCAGCTCGGCGGCGGTATCGTACGTACCATCGCAATGGGTTCCTCCGACGGTCTGCGTCGCGGTCTGGATGTAAAAGACCTCGAA
      8770      8780      8790      8800      8810      8820      8830      8840      8850      8860      8870      8880

  H  P  I  E  V  P  V  G  K  A  T  L  G  R  I  M  N  V  L  G  E  P  V  D  M  K  G  E  I  G  E  E  E  R  W  A  I  H  R  A
CACCCGATTGAAGTCCCGGTAGGTAAAGCGACTCTGGGCCGTATCATGAACGTACTGGGTGAACCGGTCGACATGAAAGGCGAGATCGGTGAAGAAGAGCGTTGGGCGATTCACCGCGCA
      8890      8900      8910      8920      8930      8940      8950      8960      8970      8980      8990      9000

  A  P  S  Y  E  E  L  S  N  S  Q  E  L  L  E  T  G  I  K  V  I  D  L  M  C  P  F  A  K  G  G  K  V  G  L  F  G  G  A  G
GCACCTTCCTACGAAGAGCTGTCAAACTCTCAGGAACTGCTGGAAACCGGTATCAAAGTTATCGACCTGATGTGTCCGTTCGCTAAGGGCGGTAAAGTTGGTCTGTTCGGTGGTGCGGGT
      9010      9020      9030      9040      9050      9060      9070      9080      9090      9100      9110      9120

  V  G  K  T  V  N  M  M  E  L  I  R  N  I  A  I  E  H  S  G  Y  S  V  F  A  G  V  G  E  R  T  R  E  G  N  D  F  Y  H  E
GTAGGTAAAACCGTAAACATGATGGAGCTCATTCGTAACATCGCGATCGAGCACTCCGGTTACTCTGTGTTTGCGGGCGTAGGTGAACGTACTCGTGAGGGTAACGACTTCTACCACGAA
      9130      9140      9150      9160      9170      9180      9190      9200      9210      9220      9230      9240

  M  T  D  S  N  V  I  D  K  V  S  L  V  Y  G  Q  M  N  E  P  P  G  N  R  L  R  V  A  L  T  G  L  T  M  A  E  K  F  R  D
ATGACCGACTCCAACGTTATCGACAAAGTATCCCTGGTGTATGGCCAGATGAACGAGCCGCCGGGAAACCGTCTGCGCGTTGCTCTGACCGGTCTGACCATGGCTGAGAAATTCCGTGAC
      9250      9260      9270      9280      9290      9300      9310      9320      9330      9340      9350      9360

  E  G  R  D  V  L  L  F  V  D  N  I  Y  R  Y  T  L  A  G  T  E  V  S  A  L  L  G  R  M  P  S  A  V  G  Y  Q  P  T  L  A
GAAGGTCGTGACGTTCTGCTGTTCGTTGACAACATCTATCGTTACACCCTGGCCGGTACGGAAGTATCCGCACTGCTGGGCCGTATGCCTTCAGCGGTAGGTTATCAGCCGACCCTGGCG
      9370      9380      9390      9400      9410      9420      9430      9440      9450      9460      9470      9480

  E  E  M  G  V  L  Q  E  R  I  T  S  T  K  T  G  S  I  T  S  V  Q  A  V  Y  V  P  A  D  D  L  T  D  P  S  P  A  T  T  F
GAAGAGATGGGCGTTCTGCAGGAACGTATCACCTCCACCAAAACTGGTTCTATCACCTCCGTACAGGCAGTATACGTACCTGCGGATGACTTGACTGACCCGTCTCCGGCAACCACCTTT
      9490      9500      9510      9520      9530      9540      9550      9560      9570      9580      9590      9600

  A  H  L  D  A  T  V  V  L  S  R  Q  I  A  S  L  G  I  Y  P  A  V  D  P  L  D  S  T  S  R  Q  L  D  P  L  V  V  G  Q  E
GCGCACCTTGACGCAACCGTGGTACTGAGCCGTCAGATCGCGTCTCTGGGTATCTACCCGGCCGTTGACCCGCTGGACTCCACCAGCCGTCAGCTGGACCCGCTGGTGGTTGGTCAGGA
      9610      9620      9630      9640      9650      9660      9670      9680      9690      9700      9710      9720
```

H Y D T A R G V Q S I L Q R Y Q E L K D I I A I L G M D E L S E E D K L V V A R
CACTACGACACCGCGCGTGGCGTTCAGTCCATCCTGCAACGTTATCAGGAACTGAAAGACATCATCGCCATCCTGGGTATGGATGAACTGTCTGAAGAAGACAAACTGGTGGTAGCGCGT
9730 9740 9750 9760 9770 9780 9790 9800 9810 9820 9830 9840

A R K I Q R F L S Q P F F V A E V F T G S P G K Y V S L K D T I R G F K G I M E
GCTCGTAAGATCCAGCGCTTCCTGTCCCAGCCGTTCTTCGTGGCAGAAGTATTCACCGGTTCTCCGGGTAAATACGTCTCCCTGAAAGACACCATCCGTGGCTTTAAAGGCATCATGGAA
9850 9860 9870 9880 9890 9900 9910 9920 9930 9940 9950 9960

unc C (ε)

G E Y D H L P E Q A F Y M V G S I E E A V E K A K K L * M A M T Y H
GGCGAATACGATCACCTGCCGGAGCAGGCGTTCTACATGGTCGGTTCCATCGAAGAAGCTGTGGAAAAAGCCAAAAAACTTTAACGCCTTAATCGGAGGGTGATATGGCAATGACTTACC
9970 9980 9990 10000 10010 10020 10030 10040 10050 10060 10070 10080

L D V V S A E Q Q M F S G L V E K I Q V T G S E G E L G I Y P G H A P L L T A I
ACCTGGACGTCGTCAGCGCAGAGCAACAAATGTTCTCTGGTCTGGTCGAGAAAATCCAGGTAACGGGTAGCGAAGGTGAACTGGGGATCTACCCTGGCCACGCACCGCTGCTCACCGCCA
10090 10100 10110 10120 10130 10140 10150 10160 10170 10180 10190 10200

K P G M I R I V K Q H G H E E F I Y L S G G I L E V Q P G N V T V L A D T A I R
TTAAGCCTGGTATGATTCGCATCGTGAAACAGCACGGTCACGAAGAGTTTATCTATCTGTCTGGCGGCATTCTTGAAGTGCAGCCTGGCAACGTGACCGTTCTGGCCGACACCGCAATTC
10210 10220 10230 10240 10250 10260 10270 10280 10290 10300 10310 10320

G Q D L D E A R A M E A K R K A E E H I S S S H G D V D Y A Q A S A E L A K A I
GCGGCCAGGATCTCGACGAAGCGCGAGCCATGGAAGCGAAACGTAAGGCTGAAGAGCACATTAGCAGCTCTCACGGCGACGTAGATTACGCTCAGGCGTCTGCGGAACTGGCCAAAGCGA
10330 10340 10350 10360 10370 10380 10390 10400 10410 10420 10430 10440

A Q L R V I E L T K K A M *
TCGCGCAGCTGCGCGTTATCGAGTTGACCAAAAAAGCGATGTAACACCGGCTTGAAAAGCACAAAAGCCAGTCTGGAAACAGGCTGGCTTTTTTTTGCGCGTGTGACCCGTCCTGAATAG
10450 10460 10470 10480 10490 10500 10510 10520 10530 10540 10550 10560

CGTTCACATAGATCCTGCTGATATAAACCCCCCTGTTTTCCTGTTTATTCATTGATCGAAATAAGAGCAAAAACATCCACCTGACGCTTAAATTAAGGTACTGCCTTAATTTTCTGCAG
10570 10580 10590 10600 10610 10620 10630 10640 10650 10660 10670 10680

ACAAAAGGCGTGACGATGGTCGAAAATGGCGCTTTCGTCAGCGGGGATAATCCGTTATTGAACAATTTATCCTCTGTCCATTTCACGATGAAAAAAATGTAGTTTTTTCAAGGTGAAGCG
10690 10700 10710 10720 10730 10740 10750 10760 10770 10780 10790 10800

Eco urf 1

M L N N A M S V V I L A A G K G T R M Y S D L P K V L H
GTTTAAATTCGTTCTCAAATTACAGTCAGGACGCGTATGTTGAATAATGCTATGAGCGTAGTGATCCTTGCCGCAGGCAAAGGCACGCGCATGTATTCCGATCTTCCGAAAGTGCTGCAT
10810 10820 10830 10840 10850 10860 10870 10880 10890 10900 10910 10920

T L A G K A M V Q H V I D A A N E L G A A H V H L V Y G H G G D L L K Q A L K D
ACCCTTGCCGGGAAAGCGATGGTTCAGCATGTCATTGATGCTGCGAATGAATTAGGCGCAGCGCACGTTCACCTGGTGTACGGTCACGGCGGCGATCTGCTAAAACAGGCGCTGAAAGAC
10930 10940 10950 10960 10970 10980 10990 11000 11010 11020 11030 11040

D N L N W V L Q A E Q L G T G H A M Q Q A A P F F A D D E D I L M L Y G D V P L
GACAACCTTAACTGGGTGCTTCAGGCAGAGCAGCTGGGTACGGGTCATGCAATGCAGCAGGCCGCACCTTTCTTTGCCGATGATGAAGACATTTTAATGCTCTACGGCGACGTGCCGCTG
11050 11060 11070 11080 11090 11100 11110 11120 11130 11140 11150 11160

I S V E T L Q R L R D A K P Q G G I G L L T V K L D D P T G Y G R I T R E N G K
ATCTCTGTCGAAACACTCCAGCGTCTGCGTGATGCTAAACCGCAGGGTGGCATTGGTCTGCTGACGGTGAAACTGGATGATCCGACCGGTTATGGACGTATCACCCGTGAAAACGGCAAA
11170 11180 11190 11200 11210 11220 11230 11240 11250 11260 11270 11280

V T G I V E H K D A T D E Q R Q I Q E I N T G I L I A N G A D M K R W L A N V T
GTTACCGGCATTGTTGAGCACAAAGATGCCACCGACGAGCAGCGTCAGATTCAGGAGATCAACACCGGCATTCTGATTGCCAACGGCGCAGATATGAAACGCTGGCTGGCGAACGTGACC
11290 11300 11310 11320 11330 11340 11350 11360 11370 11380 11390 11400

N N N A Q G E Y Y I T D I I A L A Y Q E G R E I V A V H P Q R L S E V E G V N N
AACAATAATGCTCAGGGCGAATACTACATCACCGACATTATTGCGCTGGCGTATCAGGAAGGGCGTGAAATCGTCGCCGTTCATCCGCAACGTTTAAGCGAAGTAGAAGGCGTGAATAAC
11410 11420 11430 11440 11450 11460 11470 11480 11490 11500 11510 11520

R L Q L S R L E R V Y Q S E Q A E K L L L A G V M L R D P A R F D L R G T L T H
CGCCTGCAACTCTCCCGTCTGGAGCGTGTTTATCAGTCCGAACAGGCTGAAAAACTGCTGTTAGCAGGCGTTATGCTGCGCGATCCAGCGCGTTTTGATCTGCGTGGTACGCTAACTCAC
11530 11540 11550 11560 11570 11580 11590 11600 11610 11620 11630 11640

G R D V E I D T N V I I E G N V T L G H R V K I G T G C V I K N S V I G D D C E
GGGCGCGATGTTGAAATTGATACTAACGTTATCATCGAGGGCAACGTGACTCTCGGTCATCGCGTGAAAATTGGCACCGGTTGCGTGATTAAAAACAGCGTGATTGGCGATGATTGCGAA
11650 11660 11670 11680 11690 11700 11710 11720 11730 11740 11750 11760

I S P Y T V V E D A N L A A A C T I G P F A R L R P G A E L L E G A H V G N F V
ATCAGTCCGTATACCGTTGTGGAAGATGCGAATCTGGCAGCGGCCTGTACCATTGGCCCGTTTGCCCGTTTGCGTCCTGGTGCTGAGTTGCTGGAAGGTGCTCACGTCGGTAACTTCGTT
11770 11780 11790 11800 11810 11820 11830 11840 11850 11860 11870 11880

E M K K A R L G K G S K A G H L T Y L G D A E I G D N V N I G A G T I T C N Y D
GAGATGAAAAAAGCGCGTCTGGGTAAAGGCTCGAAAGCTGGTCATCTGACTTACCTGGGCGATGCGGAAATTGGCGATAACGTTAACATCGGCGCGGGAACCATTACCTGCAACTACGAT
11890 11900 11910 11920 11930 11940 11950 11960 11970 11980 11990 12000

G A N K F K T I I G D D V F V G S D T Q L V A P V T V G K G A T I A A G T T V T
GGTGCGAATAAATTTAAGACCATTATCGGCGACGATGTGTTTGTTGGTTCCGACACTCAGCTGGTGGCCCCGGTAACAGTAGGCAAAGGCGCGACCATTGCTGCGGGTACAACTGTGACG
12010 12020 12030 12040 12050 12060 12070 12080 12090 12100 12110 12120

R N V G E N A L A I S R V P Q T Q K E G W R R P V K K K *
CGTAATGTCGGCGAAAATGCATTAGCTATCAGCCGTGTGCCGCAGACTCAGAAAGAAGGCTGGCGTCGTCCGGTAAAGAAAAAGTGATTCTGGCCGGCTAACCCGGTCACATGGGATGAG
12130 12140 12150 12160 12170 12180 12190 12200 12210 12220 12230 12240

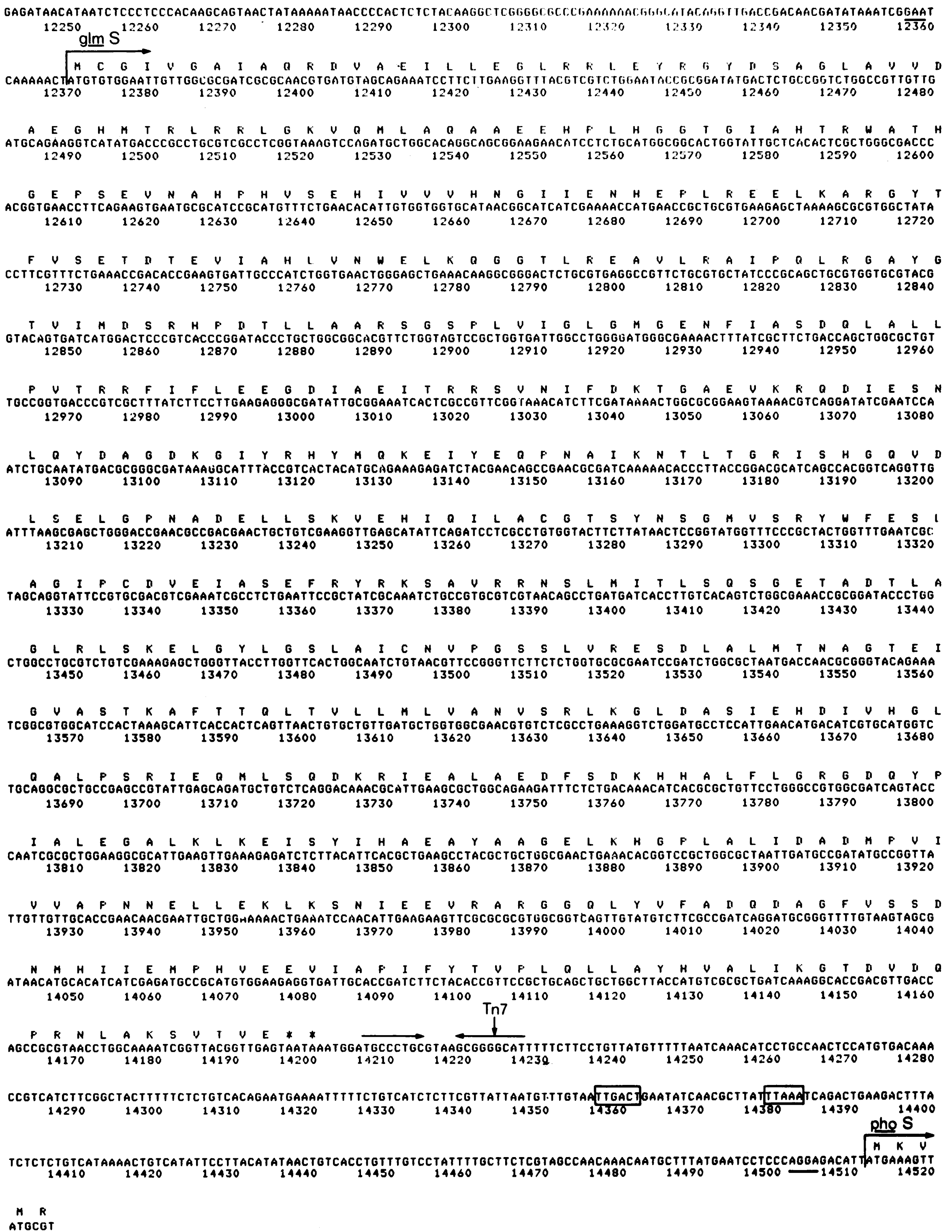


Fig. 2. *Nucleotide sequence of E. coli DNA from oriC to phoS and the amino acid sequences it codes for*
Transcription and translation are from left to right. Some of the more important control regions are marked. Each gene is marked above the proposed points for initiation of translation. Proposed ribosome-binding sites (Shine and Dalgarno sequences) are underlined. Promoter sequences are boxed and labelled −10 and −35. The point of initia-

long synthetic primer complementary to a region adjacent to linker sequence of M13mp7 (Duckworth *et al.*, 1981). In the final stages of the work the modifications described by Biggin *et al.* (1983) were employed. Sequences were compiled and analysed using computer programs described by Staden (1982*a*,*b*, 1984).

Transcriptional studies

E. coli RNA polymerase was a gift from Dr. A. A. Travers. For the preparation of run-off transcripts *in vitro* the following conditions were employed (Travers, 1981). Reactions contained the following ingredients: 0.04M-Tris/HCl (pH 7.9), 0.01M-$MgCl_2$, 0.075M-KCl, 0.6mM-2-mercaptoethanol, 0.1mM-EDTA, 2mM-ATP, 1mM-GTP, 25mM-CTP, 4μM-[α-^{32}P]UTP (50mCi/mmol; Amersham International) and 5nM-DNA fragment. Reaction mixtures were preincubated for 5min at 32°C and RNA synthesis was started by addition of RNA polymerase holoenzyme (100nM). The reactions were performed at 32°C for 30min and terminated by the addition of an equal volume of formamide dye mix. The products were separated by polyacrylamide-gel electrophoresis in the presence of urea and the gels autoradiographed. For end-labelling of RNA transcripts with [γ-^{32}P]GTP or ATP (14Ci/mmol) the reaction conditions were as above except that the nucleotide concentrations were 0.05mM for ATP and GTP and 0.125mM for GTP and UTP.

Results and discussion

Determination of the DNA sequence

The DNA sequences described in Fig. 2 were determined by the dideoxy chain termination procedure (Sanger *et al.*, 1977) coupled with cloning into bacteriophage M13 (Messing & Vieira, 1982). Three different strategies were applied at various stages of the work. Initially the sequence of the 2.5kb *Eco*RI fragment, R1, and much of the sequence of the 4.5kb *Eco*RI fragment, R2, were determined with a random shotgun strategy employing restriction enzymes with 4bp recognition sequences (Sanger *et al.*, 1982). However, it proved to be difficult to isolate sufficient clones covering part of R2 (because of unfavourable distribution of suitable restriction enzyme sites). Therefore, the sequence of this region was completed by cloning fragments generated by sonication of part of R2 (represented by the 2.9kb *Pst*I–*Eco*RI fragment NP3.2; see Fig. 1). The method of generating a random set of clones has generally superseded the earlier restriction-digest shotgun strategy.

A different directed sequencing strategy was used to determine the sequence between *ori*C and the *Eco*RI site in *unc*A. Existing information about restriction enzyme sites was used to prepare a series of primary M13 subclones, M13mp7.NB4, M13mp3.NH1.5 and M13mp3.NH4, as shown in detail in Fig. 1. Sequences were determined directly from these subclones in both orientations of the DNA. These sequences were then analysed with the computer program CUTSIT (Staden, 1984) for the presence of further 6bp restriction enzyme recognition sites. The purified inserts from the primary subclones were then digested with the appropriate restriction enzyme and cloned into the appropriate M13 vector, thereby generating further simple mixtures of clones from which recombinants could be isolated to extend the pre-existing sequences, and also to provide a reverse orientation sequence of the original data. Thus, the sequence was built up in an orderly and rapid manner in both orientations, as summarized in Fig. 3. The sequence of 14.5kb presented in Fig. 2 extends from the minimal origin of replication to the beginning of *pho*S. Nucleotides 1–420 were not determined in the present work but are taken from published data (Sugimoto *et al.*, 1979; Meijer *et al.*, 1979). No formal overlap has been determined through the *Bam*HI site near base 420, but it is likely from restriction analysis and transcription/translation experiments *in vitro* that this alignment is correct (see Lother & Messer, 1981). Therefore, the sequence of bases 1–420 has been included for completeness since it contains the coding region of the first four amino acids of *gid*A and its transcriptional promoter.

The other extremity of the sequence lies within a sequence determined by Lichtenstein & Brenner (1982) around the unique site of insertion of the transposon Tn7. This sequence (as independently verified here) overlaps that of *pho*S encoding the phosphate-repressible periplasmic phosphate-

---

tion of transcription of *unc* is marked 1. Some inverted repeated sequences that could form stable hairpin loop structures are overscored with arrows. Nucleotides 1–420 (Meijer *et al.*, 1979; Sugimoto *et al.*, 1979) were not sequenced in the present work and no overlap has been determined through the *Bam*H1 site marked in the region of nucleotide 420. In the present work the sequence was determined up to base 14520 (and beyond). Surin *et al.* (1984) have described the sequence from base 14233 onwards and Magota *et al.* (1984) sequenced from base 14242 onwards. The site of insertion of transposon, Tn7, was defined by N. J. Gay, V. L. J. Tybulewicz & J. E. Walker (unpublished work).

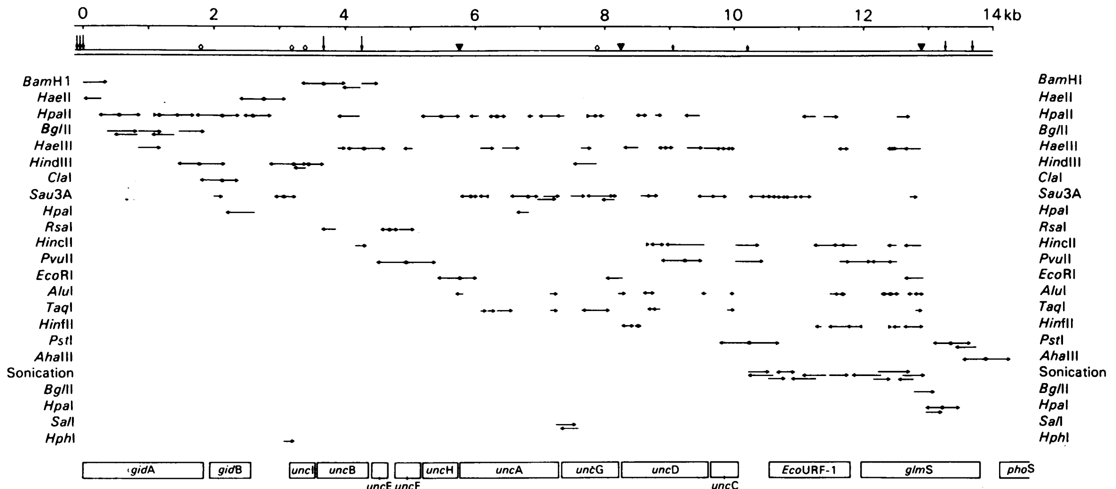


Fig. 3. *Summary of clones used to establish the sequence*

They are aligned with restriction and gene maps of the region. The scale is in kilobases. Restriction sites: ↓, *Bam*H1; ◇, *Hin*dIII; ▽, *Eco*RI; ↓, *Pst*I. Horizontal arrows represent extent of the sequences determined and their orientations.

binding protein (Magota *et al.*, 1984; Surin *et al.*, 1984).

The central portion of the sequence is taken up by the *unc* operon which has been presented previously in fragmentary form. However, the sequence presented in Fig. 2 contains two corrections of the sequences presented by Gay & Walker (1981*a*) and Saraste *et al.* (1981). The sequence of nucleotides 7076–7084 in *unc* was previously incorrectly presented as ATGCTGCAA (encoding Met-Leu-Gly) (Gay & Walker, 1981*a*). This is now corrected to CGTGCTGCA (encoding Arg-Ala-Ala). In their sequence of *unc*C, Kanazawa *et al.* (1982*b*) found an additional C residue corresponding to position 10454 in the present sequence. We have now re-examined the sequencing gels and find that a C residue is clearly present at this position. This changes the *C*-terminal amino sequence of the ε subunit of ATP synthase from Leu-Ser-Ser (Saraste *et al.*, 1981) to Val-Ile-Glu and extends it by five additional amino acids. The sequence in Fig. 2 differs in more than 40 other positions from sequences of the *unc* operon presented by Kanazawa *et al.* (1981*a,b*) and Mabuchi *et al.* (1981). These differences have been listed by Gay (1981) with detailed experimental documentation of the present sequence where these differences occur. Nucleotides 3832–6331 have also been determined independently by Nielsen *et al.* (1981) and their sequence is in complete agreement with this part of the sequence presented here.

Identification of genes

The identification of H^+-ATPase subunits from protein sequences (Sebald & Hoppe, 1982; Walker *et al.*, 1982) has been described previously.

Open reading frames (potential genes) in the rest of the sequence presented in Fig. 2 were identified with the aid of the general sequence analysis program, ANALYSEQ (Staden, 1984). One option in this program searches for potential genes by comparison of the codon usage along the sequence with the codon usage of a 'standard' *E. coli* gene. As a further help to delineating genes the program also plots the distribution of start and stop codons on the same diagram. This analysis predicts the presence of two genes (corresponding to *gid*A and *gid*B) upstream of the *unc* operon and two further genes downstream of *unc* (Fig. 4).

Protein initiation and termination sites

Table 1 lists the proposed start points for translation. With the exception of the genes for the eight subunits of ATP synthase (*unc*B–*unc*C) these have not been rigorously established; they have been chosen by finding an ATG or GTG preceded by the best 'Shine and Dalgarno' (SD) sequence (Shine & Delgarno, 1974). All of the genes except for *Eco*URF-1 (which uses TGA) terminate with a TAA codon. Tandem stop codons occur twice, in *unc*E (TAGTAA) and *glm*S (TAATAA). It has been estimated previously that approx. 13% of all *E. coli* genes contain termination signals TAG or

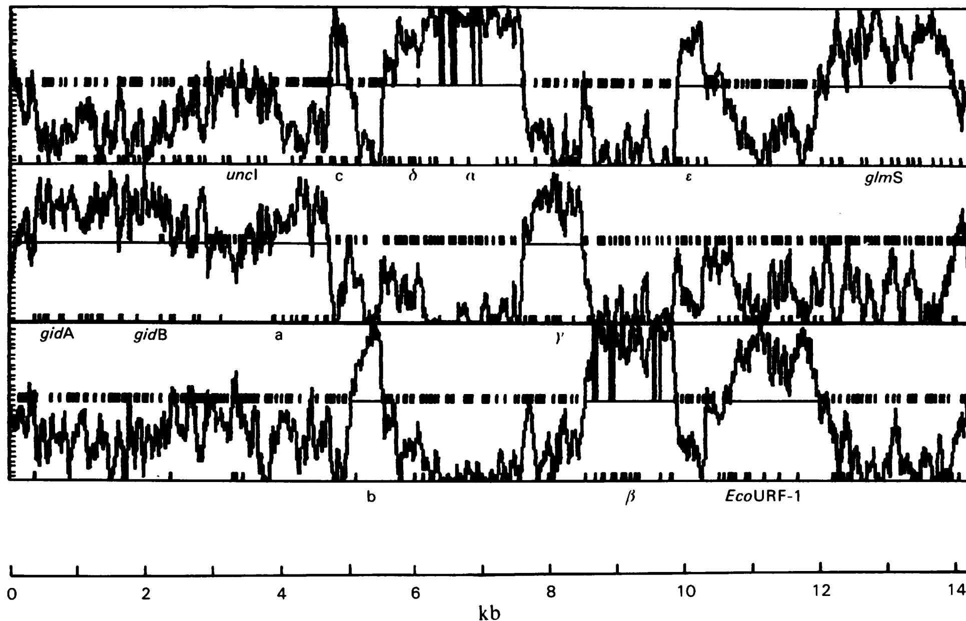


Fig. 4. *Gene predictions for the DNA in the vicinity of the unc operon by using the codon preference method of Staden & McLachlan (1982)*

This is a function of ANALYSEQ (Staden, 1984). The *x*-axis represents the DNA sequence and the probability of coding is plotted in the *y* direction. The method assumes that codon preferences for neighbouring genes are similar and hence that the codon usage of known genes from the same organism can be used as standard. In the present calculation the codon usage of the ATPase β-subunit (*unc*D) was employed for this purpose. Probabilities of coding are calculated by sliding a window of 35 codons along the sequence one codon at a time. For each position of the window, the codons in each of the three reading frames of the DNA are compared with the standard and the corresponding probabilities of coding calculated. Probabilities are plotted for each reading frame, one above the other, every five codons. A solid horizontal line at the mid point of a reading frame (at 50% probably) indicates which of the three frames is most likely to be coding. Initiation codons ATG, GTG are marked as vertical bars along the base of each plot, and termination codons as vertical bars along the 50% level.

TAA in tandem with a second nonsense triplet (Lu & Rich, 1971). It is thought that the feature ensures that readthrough ('leakiness') does not occur.

Possible secondary structure

The sequence has not been searched exhaustively for potential secondary structures. However, it was noticed that such structures were present in intergenic (and interoperonic) regions. They are indicated in Fig. 2. Two of them (following *unc*C and *glm*S) are followed by runs of T residues characteristic of rho-independent transcription termination sites. The significance of the other structures remains obscure. Brusilow *et al.* (1982) have discussed the possible role of other secondary structures in the regulation of translation of the *unc* operon.

Transcription of the genes

A notable aspect of the region of the *E. coli* chromosome from *ori*C to *pho*S is that all genes are transcribed in the same direction, away from *ori*C. The following section describes features in the sequence that may serve as transcriptional promoters. They have been identified by visual inspection of the non-coding sequences before genes for the presence of sequences resembling canonical promoters (Pribnow, 1978). Some of these features have been studied experimentally. For example, the *gid*A promoter is probably the weak promoter (indicated in Fig. 2) described by Lother & Messer (1981). The intergenic region between *gid*A and *gid*B contains a potential secondary structure of unknown significance (see Fig. 2) and canonical promoter sequences are not apparent in this region, consistent with the co-transcription of the two genes. The region following *gid*B does not evidently contain a typical rho-independent terminator, although it may be assumed from transcriptional studies of *unc* that the transcription of *gid*A and B does terminate in this region.

unc. The region preceding *unc*I contains a canonical promoter sequence from which transcription would start 73 bp upstream of *unc*I (Gay

Table 1. *Potential ribosome-binding sites (Shine and Dalgarno sequences)*
The proposed ribosome-binding sites are underlined and the initiation codons boxed.

Gene	Sequence
*gid*A	T T A T G A G G C A A T C A C C [A T G]
*gid*B	A G A A C A G G T A A T C A C C [G T G]
*unc*I	C T C G A A G G G A G C A G G A [G T G]
*unc*B	A A A A G G C A T C [A T G]
*unc*E	C T G G A G A C T G T C [A T G]
*unc*F	A G A G G C A T T G T G C T [G T G]
*unc*H	A A G G A G G G A G G G G C T G [A T G]
*unc*A	G G G G A C T G G A G C [A T G]
*unc*G	T T G A G G A G A A G C T C [A T G]
*unc*D	G T A G A G G A T T T A A G [A T G]
*unc*C	A A T C G G A G G G T G A T [A T G]
URF-1	C A G T C A G G A C G C G T [A T G]
*glm*S	C G G A A T C A A A A A C T [A T G]

& Walker, 1981*b*). In order to verify the presence of an active promoter, run-off transcripts with RNA polymerase were made from various fragments prepared from M13 replicative forms of various clones covering nucleotides 3165–6200. These experiments demonstrate that the proposed promoter is the only active promoter *in vitro* in the proximal region of the *unc* operon. Fig. 5 shows the products of transcription *in vitro* of a 240 bp *Sau*3A–*Hin*dIII fragment (SU6) (nucleotides 3395–3634) compared with that of a synthetic gene (SSU2) containing the tRNA^{Tyr} promoter (Ryan *et al.*, 1979). The product of SSU2 is known to be 142 bases in length. A single major transcript with a mobility slightly greater than the SSU2 transcript is produced by the *unc* promoter fragment. The length of this transcript predicted from the DNA sequence is 135 bases. Thus, the experimentally determined size corresponds well with the prediction.

To localize further the transcription initiation base, the transcription reactions *in vitro* were performed with [γ-^{32}P]GTP and [γ-^{32}P]ATP. Under these conditions radioactive label can only be incorporated into the initiating base. Both the SSU2 and SU6 fragments directed the synthesis of labelled products of the expected mobilities in the presence of [γ-^{32}P]GTP but not with [γ-^{32}P]ATP. The SSU2 transcript is known to initiate with G (Ryan *et al.*, 1979). This experiment confirmed the transcription initiation nucleotide proposed in Fig. 2. This conclusion has also been reached independently by Porter *et al.* (1983) and Kanazawa *et al.* (1982*a*). It is also consistent with Tn10 insertions in this region described by von Meyenburg *et al.* (1982) and DNAase footprinting experiments (Kanazawa *et al.*, 1982*a*). Other weak promoters also occur between the *unc* promoter and *unc*B, but their significance, if any, is unclear (see Porter *et al.*, 1983).

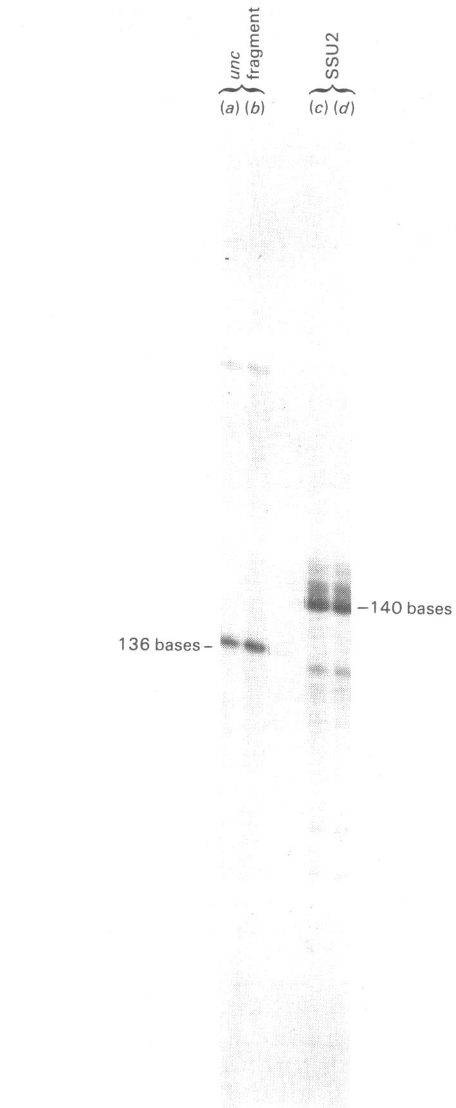


Fig. 5. *Transcription in vitro from the unc promoter*
[α-^{32}P]UTP-labelled products of transcription *in vitro* directed by (*a*, *b*) SU6 containing the *unc* promoter and (*c*, *d*) SSU2 (synthetic) *tyr*T promoter fragments.

Table 2. *Sizes of proteins encoded in the sequence*
References: (*a*), Hansen *et al.* (1981); (*b*), Brusilow *et al.* (1983).

Protein	Size: From DNA (Da)	Size: From gels (kDa)	Reference
*gid*A	69366	70	(*a*)
*gid*B	23399	25	(*a*)
*unc*I	14072	14	(*b*)
*unc*B	30304	23	(*a*)
*unc*E	8256	8.5	(*a*)
*unc*F	17232	18	(*a*)
*unc*H	19332	20	(*a*)
*unc*A	55328	58	(*a*)
*unc*G	31578	31	(*a*)
*unc*D	50326	52	(*a*)
*unc*C	15069	12	(*a*)
*eco*URF-1	49163	–	
*glm*S	66867	–	

It is also notable that in the extensive non-coding region preceding *unc*I and upstream of the *unc* promoter other conserved elements are to be found (see Fig. 2). These could conceivably act as RNA polymerase binding sites and serve to concentrate the enzyme in the vicinity of a major promoter. Such sequences have been demonstrated upstream of the *E. coli tyr*T promoter by Travers *et al.* (1983) (Travers, 1984). Some support for this proposal comes from the re-examination of the footprinting data of Kanazawa *et al.* (1982*a*). In their experiments it is apparent that protection was obtained in regions in which two of the proposed locators lie; these are the regions nucleotides 3397–3419 and 3429–3457 (see Fig. 2). Transcription of the *unc*I operon appears to terminate after *unc*C (see above). Other intergenic non-coding sequences are to be found in the operon. Before and after *unc*E are sequences homologous to each other and to the −10 region of the *trp* promoter (Pribnow, 1978) but their significance is unclear.

EcoURF-1 and glmS. A number of weak promoter sequences are indicated by ANALYSEQ upstream of *Eco*URF-1 (results not shown). However, in the absence of experimental data it is not possible to decide whether or not *Eco*URF-1 and *glm*S are co-transcribed. However, transcription of the region appears to terminate after *glm*S (see Fig. 2) before *pho*S. A potential promoter for *pho*S (see Fig. 2) has been pointed out by Surin *et al.* (1984).

The gidA and gidB proteins

The DNA sequence of *gid*A encodes a protein predicted to have a molecular mass of 69286 Da in agreement with the size determined for the product of linked transcription and translation *in vitro* of this region (see Table 2). The predicted molecular mass for the *gid*B protein is similarly in accord with earlier estimates (Table 2). Both proteins appear to have a content and distribution of hydrophilic amino acids that would suggest that they are soluble globular proteins (Kyte & Doolittle, 1982). Runs of hydrophobic amino acids that have been associated with spans buried in membranes are conspicuously absent. So it seems unlikely that either *gid*A or *gid*B has properties that would suggest that either might correspond to a component of the outer membrane binding the origin of replication. The phenotype of *gid*A and *gid*B, glucose inhibition of division, is rather vague and the biochemical function of the gene products remains unclear.

In an attempt to gain some insight, the protein sequences have been compared with those of a

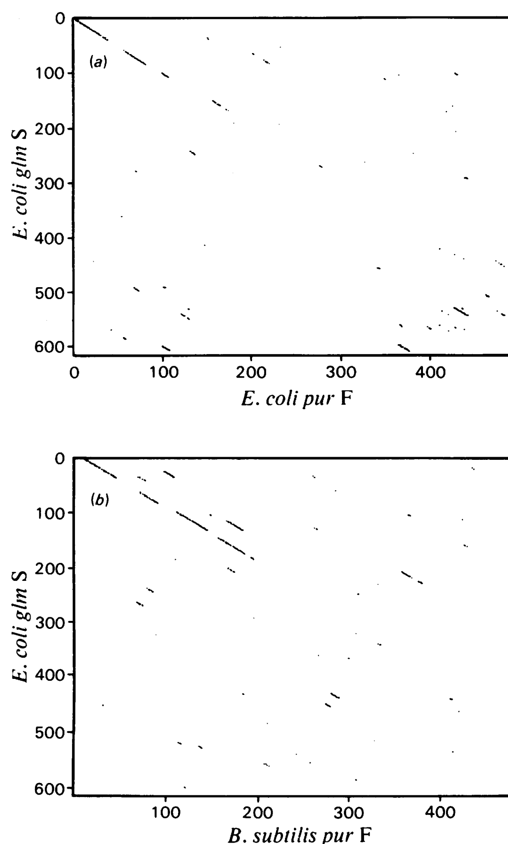


Fig. 6. *Comparison of protein sequence of glucosamine synthetase (glmS) with glutamine phosphoribosylpyrophosphate amidotransferase (pur F) from (a) E. coli and (b) B. subtilis by using DIAGON (Staden, 1982a)*
In the calculation a window of 25 and a score of 280 were employed.

Table 3. *Amino acid sequence in region of putative active site of glucosamine synthetase and other amidotransferases*
Abbreviations and references: *glmS*, glucosamine synthetase; (*a*), PRPP-AT, glutamine phosphoribosylpyrophosphate amidotransferase from *E. coli* (Tso *et al.*, 1982*a,b*); (*b*), PRPP-AT from *Bacillus subtilis* (Vollmer *et al.*, 1983; Makaroff *et al.*, 1983); (*c*), CPS, small subunit of *E. coli* carbamoyl phosphate synthetase (Piette *et al.*, 1984); (*d*) CPS from yeast (Nyunoya & Lusty, 1984); (*e*), ASII, anthranilate synthetase component II from *E. coli* (Yanofsky *et al.*, 1981); (*f*) ASII from *Serratia marcescens* (Tso *et al.*, 1980); (*g*), PABSII, *p*-aminobenzoate synthetase from *E. coli* (Kaplan & Nichols, 1983); (*h*), ASII from *Pseudomonas putida* (Kawamura *et al.*, 1978); (*i*) ASII from *Neurospora crassa* (Schechtman & Yanofsky, 1983); (*j*), FGAR-AT, formylglycinamide ribonucleotide amidotransferase from *Salmonella typhimurium* (Dawid *et al.*, 1963); (*k*), FGAR-AT from chicken (Ohnoki *et al.*, 1977). *Denotes sites of reaction with inactivating alkylating agents (see the text). Identities and conservative substitutions are boxed.

	Protein	Residues	Sequence	Reference
(*A*)	*E. coli glmS*	1-15	M C G I V G A I A Q R D V A E	This work
	E. coli PRPP-AT	1-14	*C G I V G I A G V M P V N Q	(*a*)
	B. subtilis PRPP-AT	1-14	*C G V F G I W G H E E A P Q	(*b*)
(*B*)	*E. coli* CPS	261-276	T D I P V F G I C L G H Q L L A	(*c*)
	Yeast CPS	256-271	D C I P I F G I C L G H Q L L A	(*d*)
	E. coli ASII	75-90	G K L P I I G I C L G H Q A I V	(*e*)
	S. marcescens ASII	75-90	G R L P I I G I *C L G H Q A I V	(*f*)
	E. coli PABSII	71-86	G R L P I L G V C L G H Q A M A	(*g*)
	Ps. putida ASII	71-86	G K L P I L G V *C L G H Q S I G	(*h*)
	N. crassa ASII	96-111	G K I P I F G V C M G Q Q C I F	(*i*)
	S. typhimurium FGAR-AT	-	A L G V *C	(*j*)
	Chicken FGAR-AT	-	G V *C D B C Q	(*k*)

† In glucosamine synthetase it is not known whether the initiator methionine is removed by post-translational processing, as in both species of PRPP-AT.

wide range of other proteins (Doolittle, 1981) using the rapid search technique described by Wilbur & Lipman (1983). However, no significant homologies were detected.

Identification of the glmS gene

Glucosamine synthetase (encoded in *glmS*) catalyses the formation of glucosamine 6-phosphate from fructose 6-phosphate and glutamine. Thus, it belongs to the family of amidotransferases that catalyse transfer of an amide group from glutamine to a substrate to form a new C–N bond. Three reactions in purine and two in pyrimidine nucleotide synthesis have been identified in which glutamine is a nitrogen donor. Also glutamine is a substrate for reactions leading to synthesis of anthranilate, *p*-aminobenzoate, histidine, asparagine, glutamate, glutaminyl-tRNA and nicotinamide adenine dinucleotide (Buchanan, 1973). Mutants in *glmS* are characterized by being unable to synthesize cell walls and have a requirement for exogenous glucosamine. Little is known about the chemical characteristics of the enzyme except that the glutamine analogue, 6-diazo-5-oxo-L-norleucine, is a potent inhibitor (Ghosh *et al.*, 1960). This property is also shared with other amidotransferases. In all cases in which the reaction site has been identified inhibition results from reaction of the analogue with an active site cysteine residue. In two amidotransferases, anthranilate synthase component II (Kawamura *et al.*, 1978) and glutamine : phosphoribosylpyrophosphate amidotransferase in *E. coli* and *B. subtilis* (Vollmer *et al.*, 1983; Tso *et al.*, 1982*b*) this cysteine residue has been identified in the sequence of the protein.

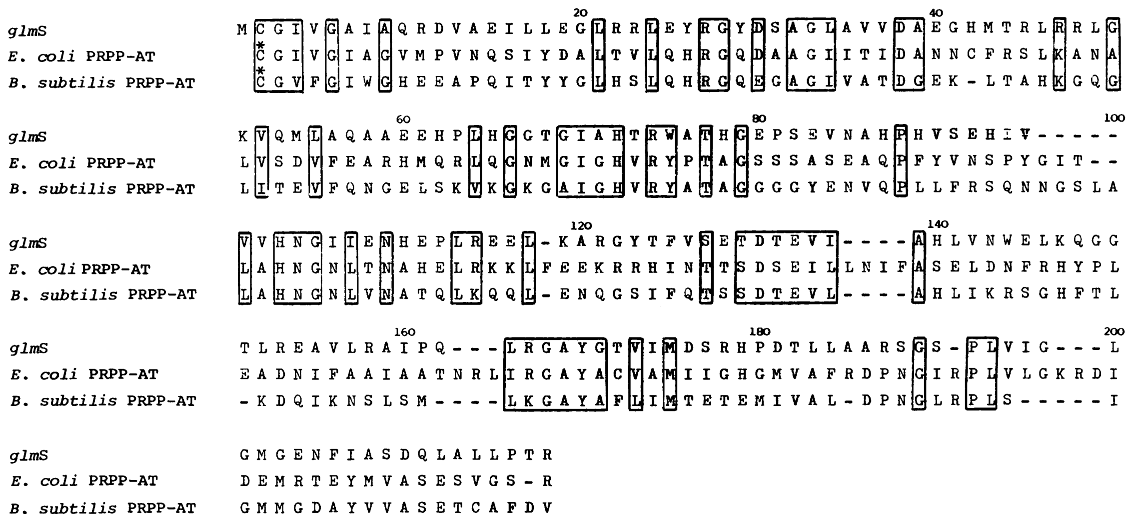


Fig. 7. *Alignment of protein sequence of glucosamine synthetase with those of glutamine phosphoribosylpyrophosphate amidotransferases (PRPP-AT) from E. coli and B. subtilis*

N-Terminal segments of the proteins are shown. They correspond to homologous regions detected with DIAGON (see Fig. 6). Conservative substitutions and identities are boxed; –, a deletion; *, sites of reaction of glutamine phosphoribosylpyrophosphate amidotransferase with 6-diazo-5-oxonorleucine (Tso *et al.*, 1982*b*; Vollmer *et al*, 1983). The *E. coli* and *B. subtilis purF* products are synthesized as precursors extended one and eleven residues respectively at their *N*-termini (Tso *et al.*, 1982*a*; Makaroff *et al.*, 1983).

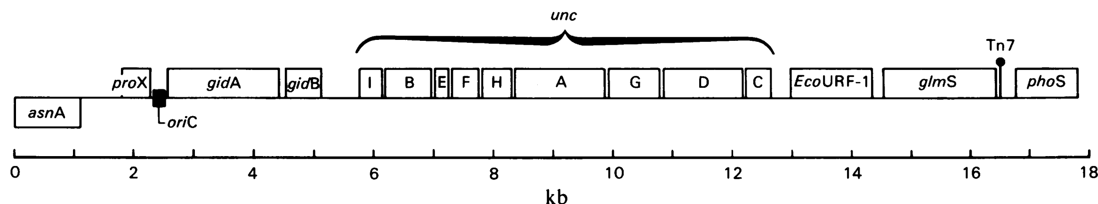


Fig. 8. *Gene map determined by DNA sequencing of the E. coli chromosome in the vicinity of the unc operon*
Genes above the line are transcribed from left to right and those below from right to left. The scale is in kb.

Other amidotransferases have also been shown to be inactivated by alkylation of a cysteine residue with other reagents. Thus, experiments have been performed to identify the sites of reaction of azaserine with *S. typhimurium* formylglycinamide:ribonucleotide amidotransferase (Dawid *et al.*, 1963) and of iodoacetic acid with the chicken enzyme (Ohnoki *et al.*, 1977), and of L-2-amino-4-oxo-5-chloropentanoic acid (another glutamine analogue) with anthranilate synthase component II (Kawamura *et al.*, 1978). From these studies and by comparison of the protein sequences of this family of proteins it appears that amidotransferases fall into two distinct groups (labelled A and B in Table 3) exemplified by carbamoyl phosphate synthetase and glutamine:phosphoribosylpyrophosphate amidotransferases (Nyunoya & Lusty, 1984). Comparison of the protein sequences predicted from the two open reading frames following the *unc* operon with these two groups demonstrates a strong homology between glutamine:phosphoribosylpyrophosphate amidotransferase and the second of the two putative *E. coli* proteins (Figs. 6 and 7), but not the first. Particularly significant is the region of homology at the *N*-terminal regions of these two proteins containing a cysteine residue. This is known to be the site of reaction of 6-diazo-5-oxonorleucine with glutamine:phosphoribosylpyrophosphate amidotransferase. Therefore, it is concluded that the second gene is that for glucosamine synthetase, *glm*S, and also that the conserved cysteine residue is likely to be the site of reaction of 6-diazo-5-oxonorleucine with this enzyme.

The first of the two reading frames following *unc*C is not evidently related to the protein sequence of any known protein (Doolittle, 1981) and its function remains unknown.

Gene organization

The organization of genes in the sector of the *E. coli* chromosome near *unc* is summarized in Fig. 8. At present it represents the most extensive sequence determined in the *E. coli* chromosome.

We thank A. A. Travers for discussion and for a gift of RNA polymerase, F. Sanger for advice and encouragement and C. J. Lusty for the preprint of a paper. M. S. was supported by an EMBO long-term Fellowship, A. N. E. by the Swiss National Foundation and N. J. G. by a Postdoctoral Fellowship. Table 3 was adapted from Nyunoya & Lusty (1984).

References

Bachman, B. J. (1983) *Microbiol. Rev.* **47**, 180–230
Biggin, M. D., Gibson, T. J. & Hong, G. F. (1983) *Proc. Natl. Acad. Sci. U.S.A.* **80**, 3963–3965
Brusilow, W. S. A., Klionsky, D. J. & Simoni, R. D. (1982) *J. Bacteriol.* **151**, 1363–1371
Brusilow, W. S. A., Porter, A. C. G. & Simoni, R. D. (1983) *J. Bacteriol.* **155**, 1265–1270
Buchanan, J. M. (1973) *Adv. Enzymol.* **39**, 91–183
Dawid, I. B., French, T. C. & Buchanan, J. M. (1963) *J. Biol. Chem.* **238**, 2178–2185
Deininger, P. L. (1983) *Anal. Biochem.* **129**, 216–222
Doolittle, R. F. (1981) *Science* **214**, 149–159
Duckworth, M. L., Gait, M. J., Goelet, P., Hong, G. F., Singh, M. & Titmas, R. C. (1981) *Nucleic Acids Res.* **9**, 1691–1706
Futai, M. & Kanazawa, H. (1980) *Curr. Top. Bioenerg.* **10**, 181–215
Gay, N. J. (1981) Ph.D. Thesis, Cambridge University
Gay, N. J. & Walker, J. E. (1981*a*) *Nucleic Acids Res.* **9**, 2187–2194
Gay, N. J. & Walker, J. E. (1981*b*) *Nucleic Acids Res.* **9**, 3919–3926
Ghosh, S., Blumenthal, H. J., Davidson, E. & Roseman, S. (1960) *J. Biol. Chem.* **235**, 1265–1273
Hansen, F. G., Nielsen, J., Riise, E. & von Meyenburg, K. (1981) *Mol. Gen. Genet.* **183**, 463–472
Kanazawa, H., Tamura, F., Mabuchi, K., Miki, T. & Futai, M. (1980) *Proc. Natl. Acad. Sci. U.S.A.* **77**, 7005–7009
Kanazawa, H., Kayano, T., Mabuchi, K. & Futai, M. (1981*a*) *Biochem. Biophys. Res. Commun.* **103**, 604–612
Kanazawa, H., Mabuchi, K., Kayano, T., Noumi, T., Sekiya, T. & Futai, M. (1981*b*) *Biochem. Biophys. Res. Commun.* **103**, 613–620
Kanazawa, H., Kayano, T., Kiyasu, T. & Futai, M. (1982*a*) *Biochem. Biophys. Res. Commun.* **105**, 1257–1264
Kanazawa, H., Mabuchi, K. & Futai, M. (1982*b*) *Biochem. Biophys. Res. Commun.* **107**, 568–575
Kaplan, J. B. & Nichols, B. P. (1983) *J. Mol. Biol.* **168**, 451–468
Kawamura, M., Keim, P. S., Goto, Y., Zalkin, H. & Heinrikson, R. L. (1978) *J. Biol. Chem.* **253**, 4659–4668
Kyte, J. & Doolittle, R. F. (1982) *J. Mol. Biol.* **157**, 105–132
Lichtenstein, C. & Brenner, S. (1982) *Nature (London)* **297**, 601–603
Lother, H. & Messer, W. (1981) *Nature (London)* **294**, 376–378
Lu, P. & Rich, A. (1971) *J. Mol. Biol.* **58**, 513–531
Mabuchi, K., Kanazawa, H., Kayano, T. & Futai, M. (1981) *Biochem. Biophys. Res. Commun.* **102**, 172–179
Magota, K., Otsuji, N., Miki, T., Horinchi, T., Tsunasawa, S., Kondon, J., Shinagawa, H. & Nakata, A. (1984) *J. Bacteriol.* **157**, 909–917
Makaroff, C. A., Zalkin, H., Switzer, R. L. & Vollmer, S. J. (1983) *J. Biol. Chem.* **258**, 10586–10593
Meijer, M., Bech, E., Hansen, H. C., Bergman, H. E. N., Messer, W., von Meyenburg, K. & Schaller, H. (1979) *Proc. Natl. Acad. Sci. U.S.A.* **76**, 580–584
Messing, J. & Vieira, J. (1982) *Gene* **19**, 269–276
Nakamura, M., Yamada, M., Hirota, Y., Sugimoto, K., Oka, A. & Takanami, M. (1981) *Nucleic Acids Res.* **9**, 4669–4676
Nielsen, J., Hansen, F. G., Hoppe, J., Friedl, P. & von Meyenburg, K. (1981) *Mol. Gen. Genet.* **184**, 33–39
Nyunoya, H. & Lusty, C. J. (1984) *J. Biol. Chem.*, **259**, 9790–9798
Ohnoki, S., Hong, B. S. & Buchanan, J. M. (1977) *Biochemistry* **16**, 1070–1076
Piette, J., Nyunoya, H., Lusty, C. J., Cunin, R., Weyens, G., Crabeel, M., Charlier, D., Glansdorff, N. & Pierard, A. (1984) *Proc. Natl. Acad. Sci. U.S.A.* **81**, 4134–4138
Porter, A. C. G., Brusilow, W. S. A. & Simoni, R. D. (1983) *J. Bacteriol.* **155**, 1271–1278
Pribnow, D. (1978) in *Biological Regulation and Development*, vol. 1 (Goldberger, R. F., ed.), pp. 219–277, Plenum Press, New York
Ryan, M. J., Ramamoorthy, B., Brown, E. L., Fritz, H. J. & Khorana, H. G. (1979) *J. Biol. Chem.* **254**, 5802–5816
Sanger, F., Coulson, A. R., Hong, G. F., Hill, D. F. & Petersen, G. B. (1982) *J. Mol. Biol.* **162**, 729–773
Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463–5467
Saraste, M., Eberle, A., Gay, N. J., Runswick, M. J. & Walker, J. E. (1981) *Nucleic Acids Res.* **9**, 5287–5296
Schechtman, M. G. & Yanofsky, C. (1983) *J. Mol. Appl. Genet.* **2**, 83–99
Sebald, W. & Hoppe, J. (1982) *Curr. Top. Bioenerg.* **12**, 1–64
Shine, J. & Dalgarno, L. (1974) *Proc. Natl. Acad. Sci. U.S.A.* **71**, 1342–1346
Staden, R. (1982*a*) *Nucleic Acids Res.* **10**, 2951–2961
Staden, R. (1982*b*) *Nucleic Acids Res.* **10**, 4731–4751
Staden, R. (1984) *Nucleic Acids Res.* **12**, 521–538
Staden, R. & McLachlan, A. D. (1982) *Nucleic Acids Res.* **10**, 141–156
Sugimoto, K., Oka, A., Sugisaki, H., Takanami, M., Nishimura, A., Yasuda, Y. & Hirota, Y. (1979) *Proc. Natl. Acad. Sci. U.S.A.* **76**, 575–579
Surin, B. P., Jans, D. A., Fimmel, A. L., Shaw, D. C., Cox, G. B. & Rosenberg, H. (1984) *J. Bacteriol.* **157**, 772–778
Travers, A. A. (1981) *J. Mol. Biol.* **141**, 91–97
Travers, A. A. (1984) *Nucleic Acids Res.* **12**, 2605–2618
Travers, A. A., Lamond, A. T., Mace, H. A F. & Berman, M. L. (1983) *Cell* **35**, 265–273

Tso, J. Y., Hermodson, M. A. & Zalkin, H. (1980) *J. Biol. Chem.* **255**, 1451–1457

Tso, J. Y., Zalkin, H., van Cleemput, M., Yanofsky, C. & Smith, J. M. (1982*a*) *J. Biol. Chem.* **257**, 3525–3531

Tso, J. Y., Hermodson, M. A. & Zalkin, H. (1982*b*) *J. Biol. Chem.* **257**, 3532–3536

Vollmer, S. J., Switzer, R. L., Hermodson, M. A., Bower, S. G. & Zalkin, H. (1983) *J. Biol. Chem.* **258**, 10582–10585

von Meyenburg, K. & Hansen, F. G. (1980) *ICN-UCLA Symp. Mol. Cell. Biol.* **19**, 137–159

von Meyenburg, K., Jørgensen, B. B., Neilsen, J. & Hansen, F. G. (1982) *Mol. Gen. Genet.* **188**, 240–248

Walker, J. E., Auffret, A. D., Carne, A., Gurnett, A., Hanisch, P., Hill, D. & Saraste, M. (1982) *Eur. J. Biochem.* **123**, 253–260

Walker, J. E. & Gay, N. J. (1983) *Methods Enzymol.* **97**, 195–218

Walker, J. E., Saraste, M. & Gay, N. J. (1984) *Biochim. Biophys. Acta Bioenerg. Rev.* **768**, 164–200

Wilbur, W. J. & Lipman, D. J. (1983) *Proc. Natl. Acad. Sci. U.S.A.* **80**, 726–730

Wolf-Watz, H. (1984) *J. Bacteriol.* **157**, 968–970

Wolf-Watz, H. & Masters, M. (1979) *J. Bacteriol.* **140**, 50–58

Wolf-Watz, H. & Norqvist, A. (1979) *J. Bacteriol.* **140**, 43–49

Wu, H. C. & Wu, T. C. (1971) *J. Bacteriol.* **105**, 455–466

Yanofsky, C., Platt, T., Crawford, I. P., Nichols, B. P., Christie, G. E., Horowitz, H., van Cleemput, M. & Wu, A. M. (1981) *Nucleic Acids Res.* **9**, 6647–6668