Volume 14 Number 10 1986 Nucleic Acids Research

Nucleotide sequence and transcriptional analysis of the *E. coli ush*A gene, encoding periplasmic UDP-sugar hydrolase (5'-nucleotidase): regulation of the *ush*A gene, and the signal sequence of its encoded protein product

Dennis M.Burns and Ifor R.Beacham

School of Science, Griffith University, Nathan, Brisbane, Queensland 4111, Australia

Received 10 February 1986; Revised and Accepted 15 April 1986

ABSTRACT

The DNA sequence of the *ush*A gene, encoding UDP-sugar hydrolase (5'-nucleotidase), has been determined. The amino-terminal sequence encodes a signal peptide whose predicted processing site is confirmed by N-terminal amino acid analysis of purified mature UshA protein. The signal sequence contains a concentration of rare codons in comparison with the mature sequence. The origins of transcription from the *ush*A promoter have been determined, using primer extension. Three transcripts, originating within a 6 bp region, were identified and might be related to three overlapping potential -10 hexamers in the *ush*A promoter region. There was a discernable change in the relative proportion of these transcripts during growth-phase regulation of the *ush*A gene.

INTRODUCTION

The *ush*A gene of *E.coli* encodes a UDP-sugar hydrolase, which also possesses 5'-nucleotidase activity (1,2). The enzyme is secreted and localised in the periplasm, and can therefore catalyse the degradation of external UDP-glucose to uridine, glucose-1-phosphate and inorganic phosphate, which can then be utilised by the cell (3).

The *ush*A gene and its protein product are of interest for several reasons: Firstly, UDP-glucose hydrolase is not only exported from the cell, but its' activity is inhibited by an intracellular protein inhibitor (1,4). The role of such inhibitors, of which there are other examples in bacteria (5), has not been established. In the case of UDP-sugar hydrolase (5'-nucleotidase), however, it has been suggested (6) that, although it is a secreted protein, a small proportion might be internally localised; the intracellular inhibitor might therefore function to inhibit potentially detrimental levels of cytoplasmic 5'-nucleotidase activity. A small pool of cytoplasmic activity is consistent with post-translational models of the secretory process (7).

Secondly, the specific activity of UDP-glucose hydrolase increases towards the end of exponential phase in batch culture (2). A similar form of 'growth-

phase regulation' has been demonstrated in the case of peptidase T in *Salmonella typhimurium* (8); in this case regulation of the *pep*T gene has been shown to be mediated by oxygen levels. However, in the case of the *ush*A gene, the mechanism remains to be elucidated.

Finally, it may be noted that the homologue of *ush*A in *Salmonella typhimurium* LT2 is a silent gene (Burns and Beacham, in preparation). Furthermore, this species contains a membrane-associated UDP-sugar hydrolase which is genetically and biochemically distinct from the *ush*A gene-product (1,9,10). The relationship of the *ush*A gene to its' silent homologue, and to the gene (*ush*B) encoding the membrane-associated UDP-sugar hydrolase in *S. typhimurium* LT2, is of considerable interest from the evolutionary point of view.

In this report we present a nucleotide sequence analysis, and a transcriptional study of the *E.coli ush*A gene in order to investigate the secretion of UDP-sugar hydrolase and the regulation of the *ush*A gene, and to facilitate future genetic studies on these and evolutionary aspects of the *ush* gene systems in *E.coli* and *Salmonella typhimurium*.

MATERIALS AND METHODS

Organisms and plasmids. pCB1 is a 18.4kb plasmid, derived from pBR322, containing a 14kb *Hin*dIII fragment of the *E.coli* genome (11). The location of the *ush*A gene on this plasmid is shown in Fig. 1. pLA7 is a 4.18 kb plasmid containing 1.81 kb of *E.coli*-derived DNA which includes the *ush*A gene (11); a complete restriction map, derived from the sequence reported herein, is also shown in Fig. 1. Plasmids were generally propagated in strain 294 (12; *end*A1, *thi*-1, *hsd*R17, *sup*E44), but for some RNA preparations were propagated in SK107 (13; *leu, rna, pnp, thr*). Growth media were as previously described (11).

Recombinant DNA techniques. General techniques for the preparation, restriction endonuclease digestion, electrophoresis and cloning of DNA, are previously described (11). Following electrophoresis, DNA fragments were isolated from agarose gels by extraction from low melting-temperature agarose (14) or by electroelution from polyacrylamide gels.

Amino-terminal amino acid sequence determination. Uridine diphospho-sugar hydrolase was purified as previously described (10). The first two N-terminal amino acids of this material were kindly determined by Dr. B. Moss (CSIRO, North Ryde) as described by Weiner et al (15).

DNA sequence determination. DNA fragments were cloned into M13 vectors and sequenced using universal primer (16) by the dideoxy chain-termination method

(17). Fragments were deleted, where necessary, with BAL31 nuclease. This involved either digestion from both ends or, alternatively, digestion of linearised pLA7 followed by re-cutting with a second enzyme. In the former case, fragments were cloned into *Sma*I cleaved M13 mp8 and their orientation determined by hybridisation using the C-test method (16); in the latter case, fragments were 'force-cloned' into appropriately cleaved M13 mp8. In neither case was it found necessary to treat the digested DNA with Klenow polymerase, prior to cloning.

Most of the sequence was derived from pLA7; however, in order to obtain substantial leader sequence it was necessary to isolate the *Hin*dIII-*Hpa*I (1.9kb) and *Hin*dIII-*Cla*I (4.69kb) fragments from pCB1 (Fig. 1). The sequence from nucleotides -195 to -93 was solely determined from these fragments, as described above.

RNA transcript mapping. The precise origin of RNA transcripts was determined using primer extension, in conjunction with dideoxy sequencing, as described by Hudson and Davidson (18): The 262 base-pairs (bp) *Bcl*I-*Pvu*II fragment, from pLA7, was cloned into M13 mp8 and single-stranded template (corresponding to the mRNA identical strand) used to prepare ^{32}P-labelled probe (18): Following the incorporation of [α-^{32}P]dCTP the reaction mix was digested with *Hpa*I, and the labelled single-stranded probe purified, following electrophoresis on a denaturing polyacrylamide gel. The probe therefore terminates with the T residue corresponding to nucleotide position 39 (see Fig. 3). Varying amounts of total RNA (1-20 µg), prepared by the method of Aiba et al (19), were hybridised with the probe, extended with reverse transcriptase (Promega Biotec), and electrophoresed on a sequencing gel. Sequencing reactions, using the same template-primer as for the probe preparation, were electrophoresed on the same gel.

RESULTS

The nucleotide sequence and protein product of the *ush*A gene.

The nucleotide sequence has been derived mainly from pLA7 (11), a detailed restriction map of which is shown in Figure 1. However 102 bp of the 5'-untranslated region were derived from the original recombinant plasmid, pCB1, from which pLA7 was later derived (see Fig. 1 and Materials and Methods). The sequencing strategy is shown in Figure 2.

The sequence, shown in Figure 3, contains an open translational reading frame of 1650bp which codes for a protein of molecular weight 60750.6; this reading frame corresponds to the direction of transcription previously

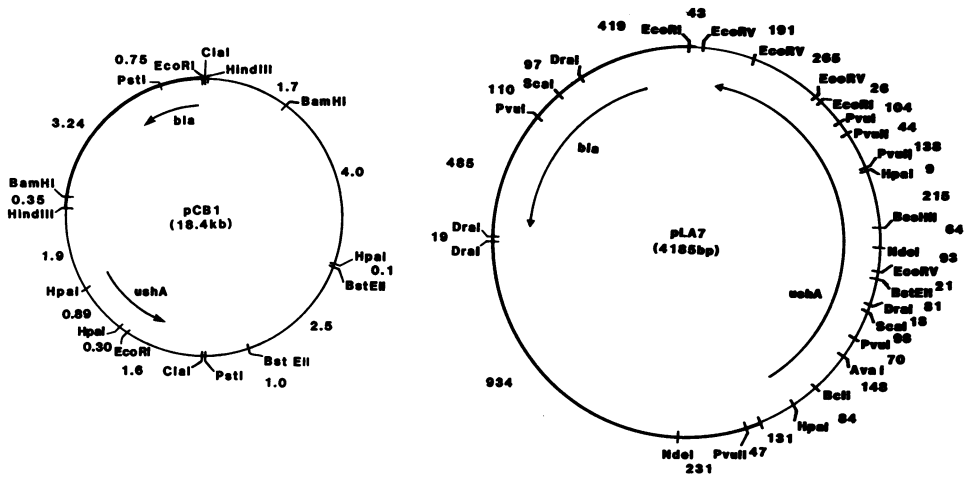

Figure 1: Restriction maps of recombinant plasmids which encode the *E.coli ushA* gene. Thick lines represent the pBR322-derived vector moiety, while thin lines represent insert DNA. The positions and direction of transcription of the *ushA* and *bla* genes are indicated. The detailed map of pLA7 was derived from the DNA sequence; fragment sizes are in bp. The distance of 131bp at the lower end of pLA7 is between the vector-insert DNA junction and the *Hpa*I site. The fragment sizes in pCB1, the parental plasmid from which pLA7 was derived, are in kb.

determined (11). No other open reading frames, on either strand, were found which would encode a protein of the reported size (M_r61,000; ref. 10).

On the basis of homology to the consensus sequences of -10 and -35 regions of *E.coli* promoters (20,21) the 5'-leader sequence of the *ushA* gene contains two putative promoters (PI and PII, Fig. 3). The -10 region for PI contains three overlapping -10 hexamers, with homology to the consensus sequence at four out of six positions in each case. We find (see below) that PI is the promoter functional *in vivo* and that there are three transcriptional start sites.

A Shine-Dalgarno sequence (see 22) is located 6 bp upstream of the ATG initiation codon (see Fig. 3). Scherer et al (23) have shown that nucleotides within the region -20 to +15 relative to the initiation codon occur non-randomly. Stormo et al (22) found that this region is limited to the nucleotides protected by the ribosome against ribonuclease when a messenger RNA is placed into an initiation complex, and inferred that some, or all, of the non-random (preferred) nucleotides may be involved in translational initiation. The sequence of the *ushA* gene, in the vicinity of

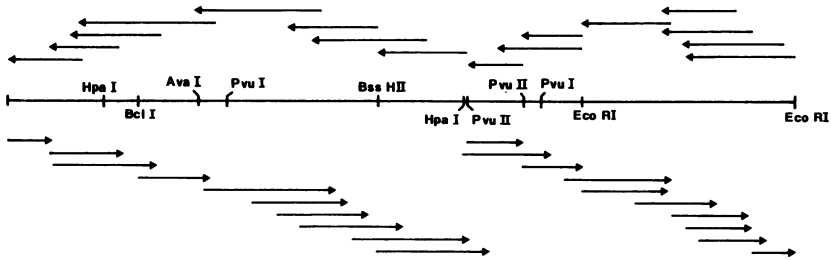

Figure 2: Sequencing strategy for the insert DNA of pLA7. Key restriction sites are shown for reference. The arrows indicate the direction of sequencing, and their lengths correspond to accurate sequence readings.

the initiation codon, conforms with these predictions in 9 out of 13 instances (Table 1). Although not of predictive value, this correlation would seem to be consistent with a functional role of these bases in translational initiation (22).

In the 3' -untranslated region of the *ushA* sequence, a GC-rich 10bp inverted repeat is present (see Fig. 3), which may function as a factor - independent transcription termination signal (see 24). This sequence corresponds to a very stable hairpin structure in the mRNA, with the ΔG value of -23.8kcal, calculated according to Tinoco (25). The consecutive T residues immediately following the hairpin loop, typical of many (see 24) but not all transcription terminators (24,26,27), are not present in the *ushA* gene. A second potential hairpin structure of moderate stability (ΔG, -7.2kcal) is, however, present 3' to the major inverted repeat (Fig. 3). Also present is a sequence (TTTG at position 1717; see Fig. 3) with a close resemblance to the TCTG consensus sequence which is observed just downstream of the termination point in many *E.coli* factor-independent terminators (24).

Codon usage in the *ushA* gene is given in Table 2. The non-random usage of synonymous codons is generally in accord with the 'rules' of Ikemura and Ozeki (28). Rare codons (see legend to Fig. 4) are used to the extent of 9.5% (52 rare codons out of a total of 550 codons). It was noticeable, however, that the signal sequence (encoding the signal peptide) contained a concentration of rare codons and we have attempted to document this in Figure 4. This figure shows the number of rare codons in a sliding window of twenty amino acid residues. Although 'windows' throughout the protein have up to five such codons, but usually less, the first twenty residues, which includes most of the signal peptide, contains eight rare codons.

```
          -181                -166                -151                -136                -121
GCGATGTAAAGAATG ATCTTATTTGTGATT ATTACCAGACTAACA TACCTGTATGCGTCC GTCTGAAGGAAGTCT

          -106                -91                 -76                 -61                 -46
CAACGCCGAATAAAA AATTTCTAATCTGGA TGCAGATTTATCTTC ACCGGACGCAGACTT GTCTATGATGTCGCG
                                                -35(PII)            -35(PI) -10(PII)
          -31                 -16          SD     -1
TCATACTATTTTTCA ACACGTTGAAATCAG GTCAGGGAGAGAAGT
    -10(PI)      15                  30                  45                  60
ATG AAA TTA TTG CAG CGG GGC GTG GCG TTA GCG CTG TTA ACC ACA TTT ACA CTG GCG AGT
MET Lys Leu Leu Gln Arg Gly Val Ala Leu Ala Leu Leu Thr Thr Phe Thr Leu Ala Ser

                 75↓                  90                  105                 120
GAA ACT GCT CTG GCG TAT GAG CAG GAT AAA ACC TAC AAA ATT ACA GTT CTG CAT ACC AAT
Glu Thr Ala Leu Ala Tyr Glu Gln Asp Lys Thr Tyr Lys Ile Thr Val Leu His Thr Asn

                 135                 150                 165                 180
GAT CAT CAT GGG CAT TTT TGG CGC AAT GAA TAT GGC GAA TAT GGT CTG GCG GCG CAA AAA
Asp His His Gly His Phe Trp Arg Asn Glu Tyr Gly Glu Tyr Gly Leu Ala Ala Gln Lys

                 195                 210                 225                 240
ACG CTG GTG GAT GGT ATC CGC AAA GAG GTT GCG GCT GAA GGC GGT AGC GTG CTG CTA CTT
Thr Leu Val Asp Gly Ile Arg Lys Glu Val Ala Ala Glu Gly Gly Ser Val Leu Leu Leu

                 255                 270                 285                 300
TCC GGT GGC GAC ATT AAC ACT GGC GTG CCC GAG TCT GAC TTA CAG GAT GCC GAA CCT GAT
Ser Gly Gly Asp Ile Asn Thr Gly Val Pro Glu Ser Asp Leu Gln Asp Ala Glu Pro Asp

                 315                 330                 345                 360
TTT CGC GGT ATG AAT CTG GTG GGC TAT GAC GCG ATG GCG ATC GGT AAT CAT GAA TTT GAT
Phe Arg Gly MET Asn Leu Val Gly Tyr Asp Ala MET Ala Ile Gly Asn His Glu Phe Asp

                 375                 390                 405                 420
AAT CCG CTC ACC GTA TTA CGC CAG CAG GAA AAG TGG GCC AAG TTC CCG TTG CTT TCC GCG
Asn Pro Leu Thr Val Leu Arg Gln Gln Glu Lys Trp Ala Lys Phe Pro Leu Leu Ser Ala

                 435                 450                 465                 480
AAT ATC TAC CAG AAA AGT ACT GGC GAG CGC CTG TTT AAA CCG TGG GCG CTG TTT AAG CGT
Asn Ile Tyr Gln Lys Ser Thr Gly Glu Arg Leu Phe Lys Pro Trp Ala Leu Phe Lys Arg

                 495                 510                 525                 540
CAG GAT CTG AAA ATT GCC GTT ATT GGG CTG ACA ACC GAT GAC ACA GCA AAA ATT GGT AAC
Gln Asp Leu Lys Ile Ala Val Ile Gly Leu Thr Thr Asp Asp Thr Ala Lys Ile Gly Asn

                 555                 570                 585                 600
CCG GAA TAC TTC ACT GAT ATC GAA TTT CGT AAG CCC GCC GAT GAA GCG AAG CTG GTG ATT
Pro Glu Tyr Phe Thr Asp Ile Glu Phe Arg Lys Pro Ala Asp Glu Ala Lys Leu Val Ile

                 615                 630                 645                 660
CAG GAG CTG CAA CAG ACA GAA AAG CCA GAC ATT ATT ATC GCG GCG ACC CAT ATG GGG CAT
Gln Glu Leu Gln Gln Thr Glu Lys Pro Asp Ile Ile Ile Ala Ala Thr His MET Gly His

                 675                 690                 705                 720
TAC GAT AAT GGT GAG CAC GGC TCT AAC GCA CCG GGC GAT GTG GAG ATG GCA CGC GCG CTG
Tyr Asp Asn Gly Glu His Gly Ser Asn Ala Pro Gly Asp Val Glu MET Ala Arg Ala Leu

                 735                 750                 765                 780
CCT GCC GGA TCG CTG GCG ATG ATC GTC GGT GGT CAC TCG CAA GAT ACG GTC TGC ATG GCG
Pro Ala Gly Ser Leu Ala MET Ile Val Gly Gly His Ser Gln Asp Thr Val Cys MET Ala

                 795                 810                 825                 840
GCA GAA AAC AAA AAA CAG GTC GAT TAC GTG CCG GGT ACG CCA TGC AAA CCA GAT CAA CAA
Ala Glu Asn Lys Lys Gln Val Asp Tyr Val Pro Gly Thr Pro Cys Lys Pro Asp Gln Gln

                 855                 870                 885                 900
AAC GGC ATC TGG ATT GTG CAG GCG CAT GAG TGG GGC AAA TAC GTG GGA CGG GCT GAT TTT
Asn Gly Ile Trp Ile Val Gln Ala His Glu Trp Gly Lys Tyr Val Gly Arg Ala Asp Phe

                 915                 930                 945                 960
GAG TTT CGT AAT GGC GAA ATG AAA ATG GTT AAC TAC CAG CTG ATT CCG GTG AAC CTG AAG
Glu Phe Arg Asn Gly Glu MET Lys MET Val Asn Tyr Gln Leu Ile Pro Val Asn Leu Lys

                 975                 990                1005                1020
AAG AAA GTG ACC TGG GAA GAC GGG AAA AGC GAG CGC GTG CTT TAC ACT CCT GAA ATC GCT
Lys Lys Val Thr Trp Glu Asp Gly Lys Ser Glu Arg Val Leu Tyr Thr Pro Glu Ile Ala
```

```
                1035                1050                1065                1080
GAA AAC CAG CAA ATG ATC TCG CTG TTA TCA CCG TTC CAG AAC AAA GGC AAA GCG CAG CTG
Glu Asn Gln Gln MET Ile Ser Leu Leu Ser Pro Phe Gln Asn Lys Gly Lys Ala Gln Leu

                1095                1110                1125                1140
GAA GTG AAA ATA GGC GAA ACC AAT GGT CGT CTG GAA GGC GAT CGT GAC AAA GTG CGT TTT
Glu Val Lys Ile Gly Glu Thr Asn Gly Arg Leu Glu Gly Asp Arg Asp Lys Val Arg Phe

                1155                1170                1185                1200
GTA CAG ACC AAT ATG GGG CGG TTG ATT CTG GCA GCC CAA ATG GAT CGC ACT GGT GCC GAC
Val Gln Thr Asn MET Gly Arg Leu Ile Leu Ala Ala Gln MET Asp Arg Thr Gly Ala Asp

                1215                1230                1245                1260
TTT GCG GTG ATG AGC GGA GGC GGA ATT CGT GAT TCT ATC GAA GCA GGC GAT ATC AGC TAT
Phe Ala Val MET Ser Gly Gly Gly Ile Arg Asp Ser Ile Glu Ala Gly Asp Ile Ser Tyr

                1275                1290                1305                1320
AAA AAC GTG CTG AAA GTG CAG CCA TTC GGC AAT GTG GTG GTG TAT GCC GAC ATG ACC GGT
Lys Asn Val Leu Lys Val Gln Pro Phe Gly Asn Val Val Val Tyr Ala Asp MET Thr Gly

                1335                1350                1365                1380
AAA GAG GTG ATT GAT TAC CTG ACC GCC GTC GCG CAG ATG AAG CCA GAT TCA GGT GCC TAC
Lys Glu Val Ile Asp Tyr Leu Thr Ala Val Ala Gln MET Lys Pro Asp Ser Gly Ala Tyr

                1395                1410                1425                1440
CCG CAA TTT GCC AAC GTT AGC TTT GTG GCG AAA GAC GGC AAA CTG AAC GAC CTT AAA ATC
Pro Gln Phe Ala Asn Val Ser Phe Val Ala Lys Asp Gly Lys Leu Asn Asp Leu Lys Ile

                1455                1470                1485                1500
AAA GGC GAA CCG GTC GAT CCG GCG AAA ACT TAC CGT ATG GCG ACA TTA AAC TTC AAT GCC
Lys Gly Glu Pro Val Asp Pro Ala Lys Thr Tyr Arg MET Ala Thr Leu Asn Phe Asn Ala

                1515                1530                1545                1560
ACC GGC GGT GAT GGA TAT CCG CGC CTT GAT AAC AAA CCG GGC TAT GTG AAT ACC GGC TTT
Thr Gly Gly Asp Gly Tyr Pro Arg Leu Asp Asn Lys Pro Gly Tyr Val Asn Thr Gly Phe

                1575                1590                1605                1620
ATT GAT GCC GAA GTG CTG AAA GCG TAT ATC CAG AAA AGC TCG CCG CTG GAT GTG AGT GTT
Ile Asp Ala Glu Val Leu Lys Ala Tyr Ile Gln Lys Ser Ser Pro Leu Asp Val Ser Val

                1635                1650
TAT GAA CCG AAA GGT GAG GTG AGC TGG CAG TAA
Tyr Glu Pro Lys Gly Glu Val Ser Trp Gln

          1668                1683                1698                1713                1728
TCCGAAAGTGCCGGA TGTTTGCATCCGGCA CAATGCTTAATCGCG GCGGGCGATATCAGC AAATTTGGCATCGAG

          1743
GATAAGCTGTCAAAC ATGAGAATTC
```

Figure 3: Nucleotide sequence of the *E.coli ushA* gene and corresponding amino acid sequence of its gene-product. The mRNA-identical strand is shown. The sequence presented is numbered with position 1 being the first base of the coding region of the *ushA* gene; bases 5' to this initiation codon are numbered negatively. The -10 and -35 regions of the two putative promoters are indicated by horizontal bars and the notations PI and PII below the appropriate sequence. It may be noted that the -10 region for PI, found to be the only functional promoter *in vivo* (see text), contains 3 overlapping -10 hexamers (see text), which are similarly indicated. The transcriptional initiation sites (see text) are indicated by dots below the appropriate bases (at nucleotide positions -26, -32, -33). Two regions of dyad symmetry which may be involved in transcriptional termination are located at nucleotide positions 1661-1684 and 1693-1707, and are indicated by horizontal arrows above the appropriate sequences. The Shine-Dalgarno sequence is indicated by a horizontal bar and the notation SD. The vertical arrow between nucleotide positions 75 and 76 indicates the processing site of the precursor protein (see text). Nucleotide position -93 represents the junction between the cloned insert and the vector moiety in pLA7; sequence 5' to this position was obtained from pCB1 (see Materials and Methods).

TABLE 1: Predictions of nucleotide distribution in the vicinity of the initiation codon.

Position[a]	Predicted nucleotides[b]	*ush*A gene
-20	A	A
- 4	A>>T	A
- 3	A	A
- 2	A/T	G
- 1	A>T	A
3	A/G	A
4	C>A	C
5	A/T	A
6	A>>C	T
10	T>>C	T
11	A>T	G
12	A	C
13	A>T	A

[a] The first base of the translation initiation codon was designated 0, according to reference 21.

[b] See reference 21.

The signal peptide

Like other secreted proteins, the mature *ush*A gene-product is derived from a higher molecular-weight precursor by proteolytic removal of an amino-terminal signal sequence (29). The amino-terminal sequence of the mature protein was determined by N-terminal sequencing to be Tyr-Glx (see Materials and Methods). Together with the observation that the processing sites of prokaryotic exported proteins are usually preceded by an ala residue (30,31), and the observed molecular weight difference between processed and precursor forms of this protein (11), the processing site can be designated as between residues 25 and 26 (see Fig. 3). Thus the signal peptide of 25 residues has a molecular weight of 2633.8, and the mature protein a molecular weight of 58,134.8; the latter value is in good agreement with the experimentally determined value of 61,000 (10).

Primer extension mapping of *ush*A transcripts

In order to determine which of the promoters, PI or PII, are used *in vivo* and to map the precise origin(s) of transcription, we have used primer extension analysis in conjunction with dideoxy-sequencing (see Materials and

TABLE 2: Codon usage in the *E.coli ushA* gene[a]

Phe	UUU	14	Ser	UCU	3	Tyr	UAU	10	Cys	UGU	0
Phe	UUC	5	Ser	UCC	2	Tyr	UAC	11	Cys	UGC	2
Leu	UUA	7	Ser	UCA	2	*	UAA	1	*	UGA	0
Leu	UUG	3	Ser	UCG	4	*	UAG	0	Trp	UGG	7
Leu	CUU	5	Pro	CCU	3	His	CAU	8	Arg	CGU	8
Leu	CUC	1	Pro	CCC	2	His	CAC	2	Arg	CGC	9
Leu	CUA	1	Pro	CCA	5	Gln	CAA	8	Arg	CGA	0
Leu	CUG	27	Pro	CCG	15	Gln	CAG	20	Arg	CGG	3
Ile	AUU	14	Thr	ACU	7	Asn	AAU	13	Ser	AGU	3
Ile	AUC	13	Thr	ACC	13	Asn	AAC	14	Ser	AGC	7
Ile	AUA	1	Thr	ACA	7	Lys	AAA	32	Arg	AGA	0
Met	AUG	16	Thr	ACG	3	Lys	AAG	9	Arg	AGG	0
Val	GUU	6	Ala	GCU	4	Asp	GAU	27	Gly	GGU	17
Val	GUC	5	Ala	GCC	13	Asp	GAC	11	Gly	GGC	23
Val	GUA	2	Ala	GCA	6	Glu	GAA	23	Gly	GGA	5
Val	GUG	27	Ala	GCG	25	Glu	GAG	12	Gly	GGG	5

[a] Asterisks indicate termination codons.

Methods). The results show (Fig. 5A and B) that three transcripts, originating at nucleotide sequence positions -33, -32 and -26, are observed, which are consistent with the use of the proximal promoter (PI; see Fig. 3). In view of the growth-phase regulation of the *ushA* gene-product (see Introduction), we have used RNA either from exponential or stationary phase cells, and obtained qualitatively identical results (Fig. 5B, cf. lanes 4 and 2,3,5). The same transcripts are observed using RNA from either plasmid-free strains, or strains containing recombinant plasmids (pCB1 or pLA7) which incorporate the *ushA* gene (Fig. 6). The signal obtained, however, with RNA from a plasmid-containing strain, overproducing UDP-sugar hydrolase, is approximately 20-fold greater than that obtained when an equivalent amount of RNA from a plasmid-free strain is used (Fig. 6); similarly, with varying amounts of RNA in the hybridisation, a correspondingly varying amount of the extended hybrid is obtained (Fig. 7A and B). These determinations are hence

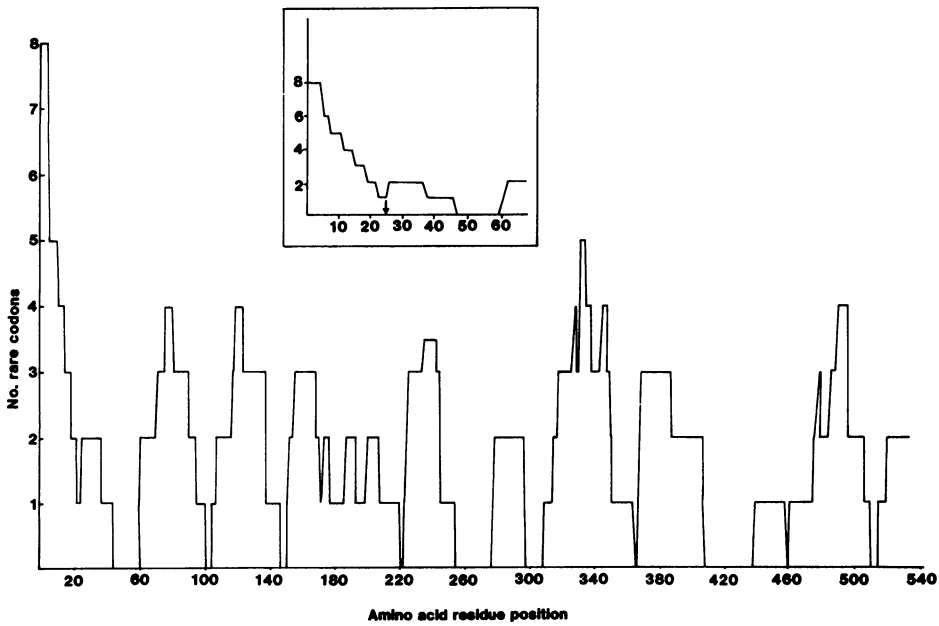

Figure 4: Distribution of rare codons in the *ushA* gene.
Based on observations of their frequency of occurrence in 25 non-regulatory genes, and relative tRNA content, we define rare codons (36) as those whose corresponding tRNA species occur with an abundancy of 0.3 or less on a scale of ~ 0-1.0 (see ref. 28), and where their percentage use [defined as the number of times a codon is used in the 25 genes referred to above, divided by the number of times all of the codons specifying the same amino acid were used] is approximately 10% or less. One exception to this definition is the arg CGA codon; this codon corresponds to an abundant tRNA species (28) but it is inefficiently recognized by the ICG-containing isoaccepting tRNA. Thus, these rare codons are: Leu: UUA, UUG, CUU, CUC, CUA; Ile: AUA; Ser: UCA, UCG, AGU; Pro: CCU, CCC; Thr: ACA; Arg: CGG, AGA, AGG, CGA; Gly: GGA, GGG.
In this figure, the number of rare codons over a running average, or 'sliding window', of 20 amino acids is graphically represented as a function of the amino acid residue position. Each data point shows the total number of rare codons in a 20 amino acid 'window', with the first residue of this window indicated on the abscissa. The inset focusses on that portion of the *ushA* gene which encodes the 25 residue signal peptide and the first 40 residues of the mature UshA protein. The cleavage site for the signal peptidase is indicated by the arrow.

at least a semi-quantitative measure of the relative amounts of each of the three mRNA species and, on this basis, their relative amounts are approximately equal, in all the above experiments; however, the longest transcript, initiating at nucleotide position -33 (Fig. 3) corresponding to the A residue (Fig. 5), is more abundant in RNA from stationary phase cells

(Fig. 5B, lane 4; Fig. 7B, lanes 1 and 2), and the shortest transcript, initiating at nucleotide position -26 (Fig. 3) corresponding to the C residue, is more abundant in RNA from exponential phase cells (Fig. 5B, lanes 2 and 3; Fig. 7A, lanes 2 and 3).

DISCUSSION

We have previously reported (11) that a less abundant protein, of molecular weight 43,000, is derived from the *ushA* gene. It is made in the form of a precursor, like the major UshA protein, and is subsequently processed (11). Thus in order to account for the lower molecular weight of this minor protein it would most likely initiate at the same position as the major *ushA* protein, but terminate at a position in the vicinity of the first *Eco*RI site downstream (nucleotide position 1223, Fig. 3). It is always observed in maxicells, and the incorporated ^{35}S-methionine cannot be chased from the (mature) protein (11). This minor protein is hence unlikely to be a biosynthetic intermediate, as in the case of some similar reported instances (32,33), but may resemble the situation with OmpA-related polypeptides, detected by immunoprecipitation (34). We have considered translational pausing as a likely explanation for the origin of the minor protein, and have searched the sequence for consecutive rare codons and for secondary structure in the mRNA. The unique occurrence of three consecutive rare codons (gly GGG, arg CGG, leu TTG) located at nucleotide positions 1156-1164 (Fig. 3), may play a role in the origin of the minor UshA protein. Indeed, it has been shown (35) that the insertion of four consecutive extremely rare codons (arg AGG) into the *E.coli cat* gene markedly reduces the level of expression, although there was no evidence presented for the appearance of a 'minor' protein product. However, because the presence of rare arginine codons reduces the expression of the Cat protein to such a large extent (see 35), the detection of a 'minor' protein product may not have been possible under the experimental conditions used. In the case of the *ushA* gene, the three codons are not extremely rare which is consistent with the observation that the 'major' UshA protein is not weakly expressed.

Although computer analysis, using the rules of Tinoco (25) for calculation of thermodynamic stability, reveals potential hairpin loops in the region between nucleotides 1065 and 1365, such structures are not unique to this region, and neither are they more stable than elsewhere in the sequence. We therefore doubt whether secondary structure in the mRNA can account for the minor UshA protein.

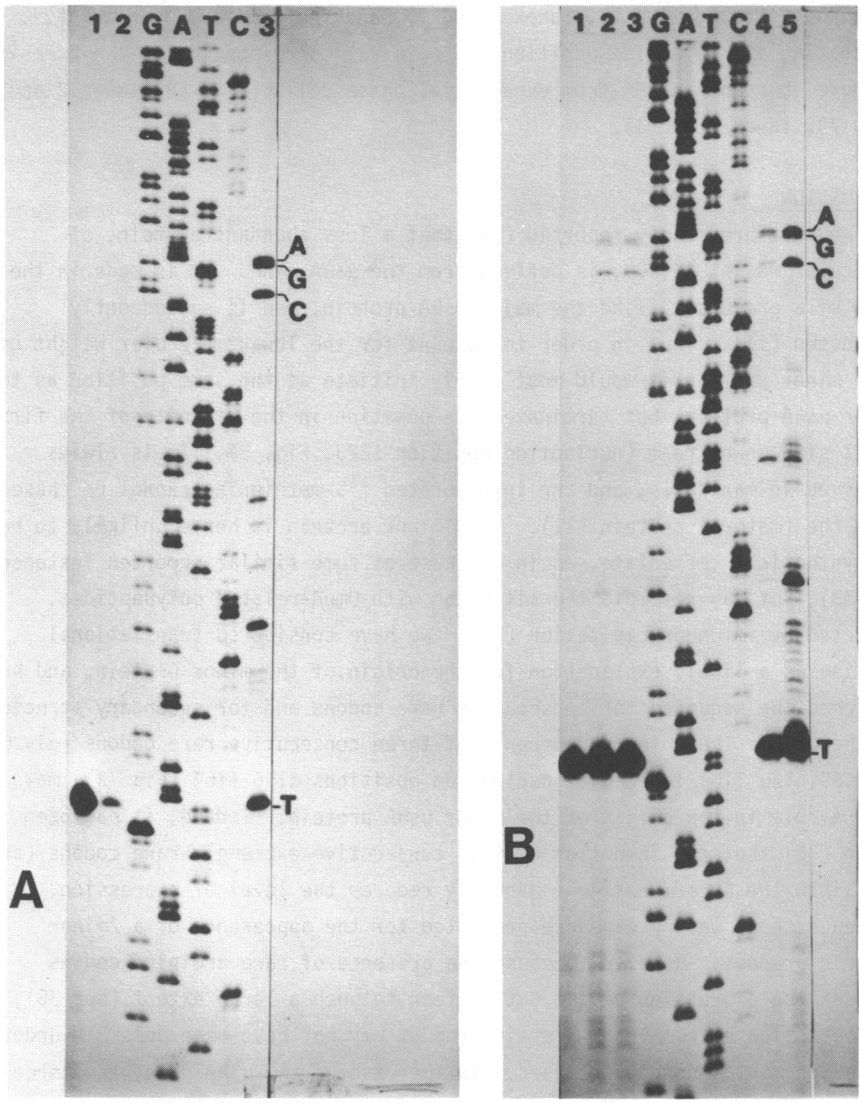

Figure 5: RNA transcript mapping by primer extension analysis. The probe used (see Materials and Methods) terminates with a T residue (indicated on Fig. 5, A and B) corresponding to the A residue at nucleotide position 39 (Fig. 3). To determine the precise origin of the RNA transcripts, the extension reactions were co-electrophoresed together with the sequencing reactions of the *Bcl*I-*Pvu*II clone from which the probe was constructed (see Materials and Methods). The base at the 3'- end of each cDNA is indicated. Intermediate bands are presumed to result from premature termination of reverse transcription. This occurs to varying degrees: Cf. lanes 2, 3 and 4, 5, Fig. 5B.

A. Lane 1: 7000 cpm of the probe extended in the absence of RNA; lane 2: 600 cpm of the probe extended in the absence of RNA; lane 3: 7000 cpm of the probe extended after annealing with 10 µg RNA from exponential phase cells containing pLA7.
B. Lane 1: The probe extended in the absence of RNA. Lanes 2-5 show the probe extended after annealing with RNA prepared from pLA7-containing cells: 2 µg RNA from exponential phase cells (lane 2); 10 µg RNA from exponential phase cells (lane 3); 2 µg RNA from stationary phase cells (lane 4); 20 µg RNA from exponential phase cells (lane 5).

It may also be noted that a search of the opposite strand revealed no open translational reading frame of sufficient length to specify a protein with the size of the minor *ush*A gene-product.

The signal peptide of the UshA protein (Fig. 8) shows structural features common to other prokaryotic signal sequences: It has two positively-charged amino acids (lys and arg) in the NH_2-terminal region, followed by a 14 amino acid stretch of uncharged, mainly hydrophobic, residues. By defining the hydrophobic 'core' as those residues encompassed between the last charged residue of the NH_2-terminus and the first charged residue towards the COOH-terminus or the last residue of the signal sequence, von Heijne (30) has analysed a number of signal sequences in terms of gross amino acid composition and hydrophobicity, and has presented a method for calculating the mean hydrophobicity per residue for each of these sequences. He found that, for ten prokaryotic signals studied, the mean hydrophobicity per residue was -3.9 kJ mol^{-1}. For the UshA signal sequence, this value is calculated to be -4.14 kJ mol^{-1}.

Other common structural features found in the UshA signal peptide are the presence of a threonine residue following the hydrophobic core, located close to the COOH-terminus of the signal peptide, and the presence of an alanine residue at the processing site which is the case for almost all known *E.coli* and *Salmonella* signal sequences; lipoprotein, which is processed by a distinct peptidase, and the minor M13 coat protein are exceptions. It is also notable, upon inspection of *E.coli* and *Salmonella* signal peptides (see 36 for references), that charged residues are almost invariably found up to five residues downstream of the cleavage site; the *ush*A protein is no exception with charged residues at positions +2, +4 and +5 (Fig. 8), where the cleavage site is defined as being between residues -1 and +1.

In regard to processing of the signal sequence, von Heijne (31) has recently carried out a statistical analysis of 20 prokaryotic and 65 eukaryotic signal sequences and found strong preference for particular amino acids around this cleavage site. His observations led to the proposal that an

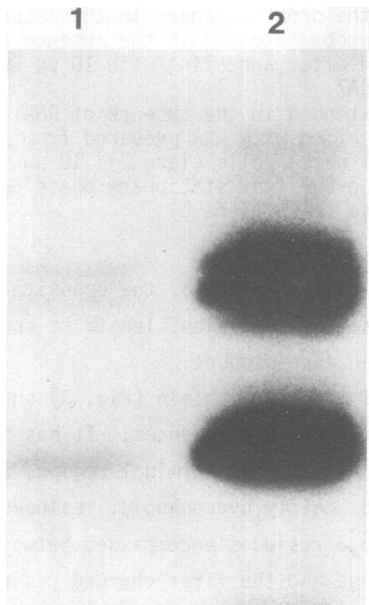

Figure 6: Comparison of cDNAs corresponding to *ushA* transcripts in plasmid-free and pCB1-containing cells by primer extension analysis. The probe used was that described in the legend to Figure 5. Only the extended products are shown. Lane 1: the probe extended after annealing with 20 µg RNA from exponential phase *ushA*$^+$ plasmid-free cells; lane 2: the probe extended after annealing with 20 µg RNA from exponential phase pCB1-containing cells.

'acceptable' cleavage site must fulfil a '(-3,-1) rule', where the cleavage site is defined as above: It must have certain preferred amino acids in these positions. Von Heijne has also suggested (31) that cleavage specificity is determined by a region which extends 5 residues downstream of the COOH-terminus of the hydrophobic core (usually position -2) to 11 residues downstream of this core (usually position +5).

In the case of the UshA protein the presence of alanine residues at positions -1 and -3 of the determined processing site (Fig. 8) is consistent with von Heijne's '(-3,-1) rule' (see above), and the processing site is five residues downstream of the COOH-terminus of the hydrophobic core. Another potential processing site, between the ala and leu residues at positions -3 and -2 respectively (Fig. 8), does not conform to this rule since its corresponding -3 position (labelled -5 on Fig. 8) is a glutamate residue; in addition, this potential processing site is located only three residues downstream of the COOH-terminus of the hydrophobic core.

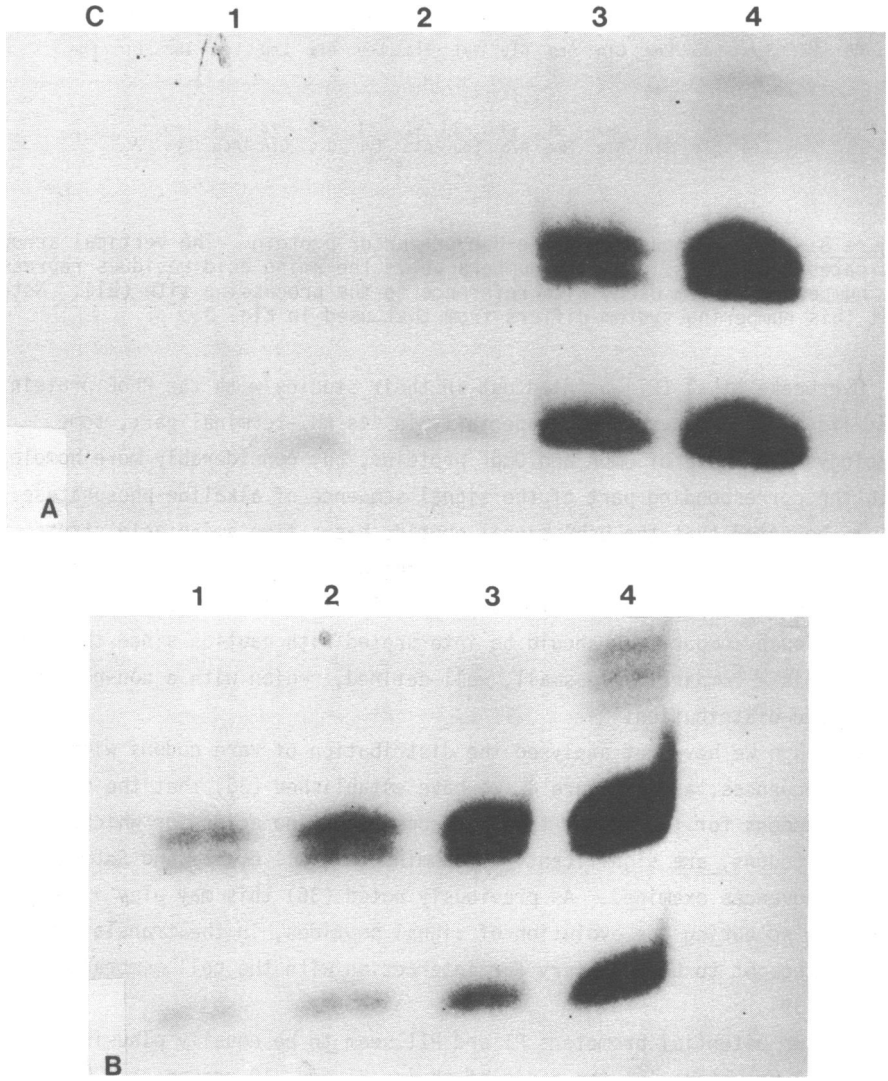

Figure 7: The effect of RNA concentration on reverse transcription. The probe (see legend to Figure 5) was extended after annealing with various amounts of RNA prepared from both exponential and stationary phase cells containing pLA7.
A. Lanes 1-4 show the extension products after annealing the probe with RNA prepared from exponential phase cells: 0.25 µl RNA (lane 1); 0.5 µl RNA (lane 2); 1.0 µl RNA (lane 3); 2.0 µl RNA (lane 4). Lane C represents the probe extended in the absence of RNA.
B. Lanes 1-4 show the extension products after annealing the probe with RNA prepared from stationary phase cells: 0.25 µl RNA (lane 1); 0.5 µl RNA (lane 2); 1.0 µl RNA (lane 3); 2.0 µl RNA (lane 4).

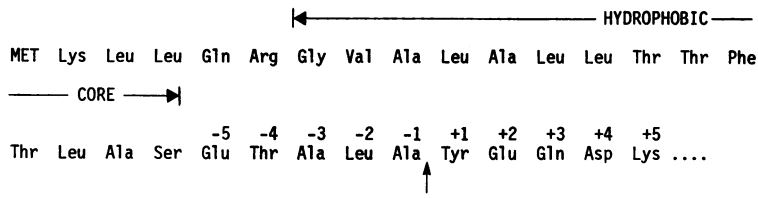

Figure 8: Signal sequence of the Ush precursor protein. The vertical arrow indicates the processing site; numbers above the amino acid residues represent the numbering system used, with reference to the processing site (31). Note that this numbering system differs from that used in Fig. 3.

Overbeeke et al (27) pointed out in their studies with the PhoE protein that its signal sequence has, especially in its NH_2-terminal part, some homology with those of OmpA and OmpF proteins, but considerably more homology with the corresponding part of the signal sequence of alkaline phosphatase. It may be noted that the UshA signal peptide has a five amino acid stretch (Ala Leu Ala Leu Leu) in the hydrophobic region in common with that of alkaline phosphatase, which differs in only 5 out of 15 nucleotide positions. However, such comparisons should be interpreted with caution since the signal sequence is a comparatively small, well-defined, region with a non-random amino acid distribution.

Although we have not analysed the distribution of rare codons within other signal sequences, as in Figure 4, we have established (36) that the occurrence of rare codons for two out of three hydrophobic amino acids, for which there are rare codons, are significantly predominant in all *E.coli* and *Salmonella* signal sequences examined. As previously noted (36) this may play a role, or have done so during the evolution of signal peptides, in the translational pausing thought to be necessary for interaction with the cell membrane (37, 38, 39).

The two potential promoters PI and PII seem to be equally plausible as functional promoters, on the basis of their -35 and -10 sequences. PI has three possible overlapping -10 hexamers, the proximal one being 17 bp from the -35 region, which is considered optimal (21). This spacing however, in the case of PII, is 15 bp which, although within the observed range of 15-20 bp (21), is not optimal. It is possibly for this reason, then, that PI is found to be the functional promoter *in vivo* (see below). However, three transcripts are observed, initiating at nucleotides -33, -32 and -26 (see Fig. 5), which might be related to the observed overlapping -10 hexamers. The

spacing between the two distal hexamers and the -35 sequence (12 and 14 bp) are not precedented (21), but, given the initial recognition of the proximal hexamer with optimal spacing, these two upstream hexamers may play a role in providing the RNA polymerase with flexibility in regard to the start site for transcription. The spacing between the initiation site as position -26 and the proximal -10 hexamer, is 7 bp; the other initiation sites are between 3-6 bp from either of the other two hexamers (see Fig. 3). The observed range is 4-8 bp, and is usually 6 or 7 bp (20, 21).

Several other instances of multiple initiation sites have been reported (see 21), although it may be noted that only rarely have the determinations been made with RNA synthesised *in vivo*, avoiding possible artifacts resulting from *in vitro* transcription. In most of these cases, as in the case reported here, at least one upstream -10 hexamer, overlapping the hexamer with 'optimal' spacing, can be discerned.

We have attempted to determine whether the growth-phase regulation of the *ush*A gene-product is related qualitatively or quantitatively to transcription of the *ush*A gene. The primer extension mapping experiments clearly show that the same three transcripts are made in the same (approximately equal) relative amounts *in vivo* in both exponential and stationary phase cells, consistent with PI (Fig. 3) being the sole functional promoter. It seems, however, that the longest transcript predominates in RNA from stationary phase cells while the shortest transcript predominates in RNA from exponential phase cells (see Results). We also attempted to determine whether the total amount of all transcripts from promoter PI is greater in RNA from stationary phase cells as compared with an equivalent amount of RNA from exponential phase cells (data not shown). Whereas in some experiments an apparently greater total quantity of *ush*A transcripts were apparent in stationary phase RNA, this was not a consistent result, and further analysis will be necessary.

Finally, it might be noted that with the sequence of the genome-derived DNA in pLA7, the complete sequence of this plasmid is now available, which should enhance its value as a direct selection vector (40). This plasmid might also be of use as a 'secretion vector': The unique *Bcl*I site is in fact located 15 codons downstream of the signal peptidase cleavage site. Cloning of appropriate sequences in the correct reading frame, at this site, should result in secretion of the resulting fusion protein.

ACKNOWLEDGEMENTS

This work was supported by the Australian Research Grants Scheme.

REFERENCES

1. Glaser, L., Melo, A. and Paul, R. (1967) J. Biol. Chem. 242, 1944-1954.
2. Neu, H.C. (1967) J. Biol. Chem. 242, 3896-3904.
3. Yagil, E. and Beacham, I.R. (1975) J. Bacteriol. 121, 401-405.
4. Neu, H.C. (1967) J. Biol. Chem. 242, 3905-3911.
5. Swartz, M.N., Kaplan, N.O. and Frech, M.E. (1956) Science 123, 50-53.
6. Taylor, N.S. and Beacham, I.R. (1975) Biochim. Biophys. Acta. 411,216-221.
7. Randall, L.L. and Hardy, S.J.S. (1984) Microbiol. Rev. 48, 290-298.
8. Strauch, K.L., Lenk, J.B., Gamble, B.L. and Miller, C.G. (1985) J. Bacteriol. 161, 673-680.
9. Osborn, M.J., Gander, J.E., Parisi, E. and Carson, J. (1972) J. Biol. Chem. 247, 3962-3972.
10. Beacham, I.R. and Wilson, M.S. (1982) Arch. Biochem. Biophys. 218, 603-608.
11. Burns, D.M., Abraham, L.J. and Beacham, I.R. (1983) Gene 25, 343-353.
12. Talmadge, K. and Gilbert, W. (1980) Gene 12, 235-241.
13. Hautala, J.A., Bassett, C.L., Giles, N.H. and Kushner, S.R. (1979) Proc. Natl. Acad. Sci. 76, 5774-5778.
14. Burns, D.M. and Beacham, I.R. (1983) Anal. Biochem. 135, 48-51.
15. Weiner, A.M., Platt, T. and Weber, K. (1972) J. Biol. Chem. 247, 3242-3251.
16. Messing, J. (1983) in Methods in Enzymology, Wu, R., Grossman, L. and Moldave, K. Eds., Vol. 101, pp. 20-78, Academic Press, New York.
17. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. 74, 5463-5467.
18. Hudson, G.S. and Davidson, B.E. (1984) J. Mol. Biol. 180, 1024-1051.
19. Aiba, H., Adhya, S. and de Crombrugghe, B. (1981) J. Biol. Chem. 256, 11905-11910.
20. Rosenberg, M. and Court, D. (1979) Ann. Rev. Genet. 13, 319-353.
21. Hawley, D.K. and McClure, W.R. (1983) Nucl. Acids Res. 11, 2237-2255.
22. Stormo, G.D., Schneider, T.D. and Gold, L.M. (1982) Nucl. Acids Res. 10, 2971-2996.
23. Scherer, G.F.E., Walkinshaw, M.D., Arnott, S. and Morre, D.J. (1980) Nucl. Acids Res. 8, 3895-3907.
24. Brendel, V. and Trifonov, E.N. (1984) Nucl. Acids Res. 12, 4411-4427.
25. Tinoco, I., Borer, P.N., Dengler, B., Levine, M.D., Uhleabeck, O.C., Crothers, D.M. and Gralla, J. (1973) Nature New Biol. 246, 40-41.
26. Smith, D.R. and Calvo, J.M. (1980) Nucl. Acids Res. 8, 2255-2274.
27. Overbeeke, N., Bergmans, H., van Mansfeld, F. and Lugtenberg, B. (1983) J. Mol. Biol. 163, 513-532.
28. Ikemura, T. and Ozeki, H. (1982) In Cold Spring Harbor Symposia on Quantitative Biology, Vol. XLVII, pp. 1087-1097, Cold Spring Harbor Laboratory, New York.
29. Silhavy, T.J., Benson, S.A. and Emr, S.D. (1983) Microbiol. Rev. 47, 313-344.
30. von Heijne, G. (1981) Eur. J. Biochem. 16, 419-422.
31. von Heijne, G. (1984) J. Mol. Biol. 173, 243-251.
32. Misra, R. and Reeves, P. (1984) Proc. Aust. Biochem. Soc. 16, 1.
33. Randall, L.L., Josefsson, L.-G. and Hardy, S.J.S. (1980) Eur. J. Biochem. 107, 375-379.
34. Crowlesmith, I. and Gamon, K. (1982) Eur. J. Biochem. 124, 577-583.
35. Robinson, M., Lilley, R., Little, S., Emtage, J.S., Yarronton, G., Stephens, P., Millican, A., Eaton, M. and Humphreys, G. (1984) Nucl. Acids Res. 12, 6663-6671.
36. Burns, D.M. and Beacham, I.R. (1985) FEBS Lett. 189, 318-324.
37. Walter, P. and Blobel, G. (1981) J. Cell. Biol. 91, 557-561.
38. Schultz, J., Silhavy, T.J., Berman, M.L., Fiil, N. and Emr, S.D. (1982) Cell 31, 227-235.
39. Hall, M.N., Gabay, J. and Schwartz, M. (1983) EMBO J. 2, 15-19.
40. Burns, D.M. and Beacham, I.R. (1984) Gene 27, 323-325.