JOURNAL OF BACTERIOLOGY, July 1985, p. 94–105
0021-9193/85/070094-12$02.00/0

Vol. 163, No. 1

Nucleotide Sequence of an *Escherichia coli* Chromosomal Hemolysin

TERESA FELMLEE, SHAHAIREEN PELLETT, AND RODNEY A. WELCH*

Department of Medical Microbiology, University of Wisconsin Medical School, Madison, Wisconsin 53706

Received 26 December 1984/Accepted 17 April 1985

We determined the DNA sequence of an 8,211-base-pair region encompassing the chromosomal hemolysin, molecularly cloned from an O4 serotype strain of *Escherichia coli*. All four hemolysin cistrons (transcriptional order, C, A, B, and D) were encoded on the same DNA strand, and their predicted molecular masses were, respectively, 19.7, 109.8, 79.9, and 54.6 kilodaltons. The identification of pSF4000-encoded polypeptides in *E. coli* minicells corroborated the assignment of the predicted polypeptides for *hlyC*, *hlyA*, and *hlyD*. However, based on the minicell results, two polypeptides appeared to be encoded on the *hlyB* region, one similar in size to the predicted molecular mass of 79.9 kilodaltons, and the other a smaller 46-kilodalton polypeptide. The four hemolysin gene displayed similar codon usage, which is atypical for *E. coli*. This reflects the low guanine-plus-cytosine content (40.2%) of the hemolysin DNA sequence and suggests the non-*E. coli* origin of the hemolysin determinant. In vitro-derived deletions of the hemolysin recombinant plasmid pSF4000 indicated that a region between 433 and 301 base pairs upstream of the putative start of *hlyC* is necessary for hemolysin synthesis. Based on the DNA sequence, a stem-loop transcription terminator-like structure (a 16-base-pair stem followed by seven uridylates) in the mRNA was predicted distal to the C-terminal end of *hlyA*. A model for the general transcriptional organization of the *E. coli* hemolysin determinant is presented.

The genetics and biochemistry of the *Escherichia coli* hemolysin have become an active area of research after recent reports of its significance in the virulence of extraintestinal *E. coli* infections (2–4, 7, 12, 43). In addition, factors influencing the level of expression of extracellular hemolysin activity are known to affect the level of homolysin-associated virulence in vivo (44). This evidence suggests that insights into the regulation of the hemolysin may be of clinical as well as academic interest. Our laboratory, along with others, has identified by transposon-mediated mutagenesis an approximate coding region of seven kilobases (kb) responsible for hemolysin synthesis (29, 30, 38, 45). Goebel and co-workers identified within this region four cistrons, *hlyC*, *hlyA*, *hlyBa*, and *hlyBb*, necessary for extracellular hemolysin activity (30, 42). There exists a close similarity in DNA sequences among *E. coli* hemolysins based on DNA-DNA hybridization (1, 19, 45). However, it has also become apparent that there are DNA sequence differences among the *E. coli* hemolysins and, not surprisingly, that these differences appear to have significant genetic, biochemical, and pathogenic consequences (12, 23, 43, 45).

Presented here is the complete DNA sequence of the hemolysin-encoding region of the recombinant plasmid pSF4000. The hemolysin determinant in this case originated from a urinary tract isolate of *E. coli* (strain J96, O4 serotype). Genetic as well as physical evidence indicates that the hemolysin was originally encoded on the J96 chromosome (15, 45).

MATERIALS AND METHODS

Bacteria and bacteriophage strains. Bacterial strains used in this study include HB101, WAF100 (HB101 transformed with pSF4000) (45) as the source of hemolysin-specific restriction endonuclease fragments, and JM101 as the host of M13 bacteriophage subclones (24). The source of hemolysin recombinant plasmids pSF4000 and pANN202-312 has been described (43, 45). DS410 (*minA minB*) was used as the host background for electrophoretic analysis of [^{35}S]methionine-labeled plasmid-encoded polypeptides present in isolated minicells (22). pUC9 and the replicative forms of the M13 vectors mp8, mp10, and mp11 (40) used for the subcloning were acquired from New England Biolabs, Inc., and P-L Biochemicals, Inc.

Media and buffers. Recipes for LB broth and YT and 5-bromo-4-chloro-3-indolyl-β-D-galactoside-YT-overlay agar media employed in growing bacteria and bacteriophage were taken from Miller (28). Tris-acetate and Tris-borate buffers utilized for DNA electrophoresis were prepared as described previously (44, 45).

Molecular cloning and DNA sequencing. The strategy for determination of the hemolysin DNA sequence by using the chain termination method (35) and M13 vectors (40) was to purify overlapping 1- to 3-kb restriction endonuclease fragments by agarose gel electrophoresis and electroelution of the DNA from the isolated gel slice. The isolated fragments were redigested with a second restriction endonuclease that is known to have multiple digestion sites within the fragment and is capable of generating ends that are insertible in the multiple cloning site of the M13 vectors. The redigested fragments were ligated with the appropriately digested M13 vector, and the chimeric phage DNA was transfected into JM101 (26). The protocols for the isolation of recombinant templates and the dideoxy-sequencing reactions were those suggested by the commercial supplier of the dideoxynucleotide mixtures, the DNA polymerase large fragment, and the restriction endonucleases (New England Biolabs). [α-^{32}P]dATP (ca. 800 Ci/mmol) was purchased from Amersham Corp. The labeled reaction mixtures were separated by electrophoresis on urea–10% acrylamide gels, the gels were dried down, and the sequence was read from X-ray film autoradiograms.

The methods used in our laboratory for the construction of in vitro derived deletions and subclones of pSF4000 and pANN202-312 have been described elsewhere (44, 45).

* Corresponding author.

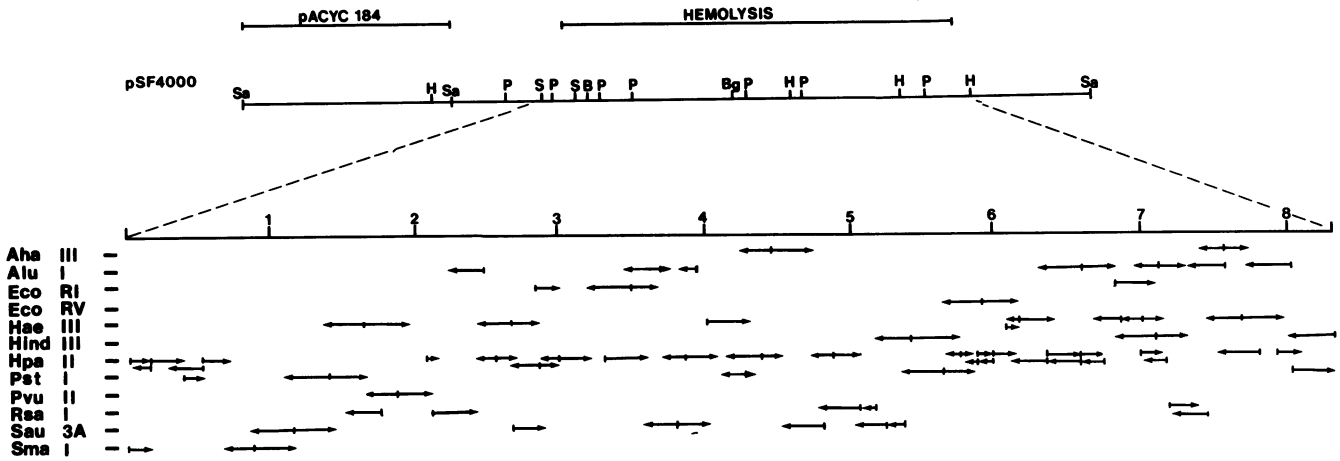


FIG. 1. Sequencing strategy of the *E. coli* hemolysin located on the recombinant plasmid pSF4000. The hemolysin was localized by Tn*1*-mediated mutagenesis to the region shown above the restriction enconuclease fragment map at the top. The 8.2-kb region that was sequenced has been expanded in the lower portion of the figure (the numbered vertical lines represent 1 kb). The direction of the sequenced DNA fragments is shown by the arrows, and the identity of the particular restriction endonuclease fragments is listed to the left. The abbreviations for the restriction endonuclease sites shown on the physical map at the top of the figure are as follows: SA, *Sal*I; H, *Hind*III; P, *Pst*I; S, *Sma*I; B, *Bam*HI; and Bg, *Bgl*II.

DNA and amino acid sequence analysis. For routine DNA sequence searches we employed the Pustell programs (33) on an IBM PC microcomputer (programs provided courtesy of International Biotechnologies, Inc.). The program for construction of codon preference plots has been described elsewhere (10). A program written to permit the drawing of hydropathy plots and the predicted polypeptide secondary structures by using the rules of Kyte and Doolittle (20) and Chou and Fasman (5) was developed by M. Gribskov and R. Burgess. These two programs were performed on a Digital VAX computer, and the plots were produced by a Hewlett-Packard 7221T plotter.

Isolation and radiolabeling of minicells. Minicells were isolated from DS410 and different DS410 plasmid transformants after overnight growth at 37°C in brain heart infusion broth. The minicells were purified by sedimentation through sucrose gradients and radiolabeled with [^{35}S]methionine (Amersham) according to the methods of Gill et al. (8). Glycerol (final concentration, 0.2%) was substituted for glucose in the minicell-labeling medium when isopropyl-β-D-thiogalactoside-induced transcription from the pUC9 *lac* promoter was being examined. Isopropyl-β-D-thiogalactoside (final concentration, 0.66 mM) was added 30 min before the addition of [^{35}S]methionine label, and incubation was continued for 90 min. Radiolabeled proteins present in sodium dodecyl sulfate-lysed minicells were separated by polyacrylamide gel electrophoresis by the method of Laemmli (21). The gels were impregnated with En^{3}Hance (New England Nuclear Corp.) according to the manufacturer's directions, and the dried gel was used for fluorography against X-Omat R film (Eastman Kodak Co.).

RESULTS

DNA sequence of the *E. coli* hemolysin. Shown in Fig. 1 is the series of overlapping M13 subclones of the hemolysin determinant present on pSF4000 used to generate the DNA sequence shown in Fig. 2. The sequence derived from these covers a continuous 8,211-base-pair (bp) region of pSF4000 within which transposon-mediated mutations and deletions indicate the hemolysin is located (45). Eighty-three percent of both DNA strands were directly sequenced, and in most instances in which the sequence for only one DNA strand was available, overlapping clones of the area helped confirm the sequence.

Association of ORFs with hemolysin cistrons. Goebel and co-workers previously established the existence of four hemolysin cistrons by use of subclones and their ability to complement a series of transposon-mediated mutations (17, 30, 42). There exist four long open-reading frames (ORFs) with ATG starts preceded by sequences resembling ribosome-binding sites (Shine-Dalgarno sequences) (36) within the sequence presented in Fig. 2. They were encoded sequentially by the DNA strand presented in the figure, and no ORFs with the aforementioned features were found on the reverse complement of the sequence presented in Fig. 2. The predicted molecular masses of the polypeptides encoded by these ORFs were, in left-to-right order (see Fig. 1), 19.7, 109.9, 79.9, and 54.6 kilodaltons (kd). Through the use of *E. coli* minicells, each putative polypeptide corresponded in mass to a similar species identified as being hemolysin-specific and physically encoded in the region presented in Fig. 2. Through the use of deletion and transposon mutants of pSF4000, as well as subcloned fragments of pSF4000 and a recombinant plasmid encoding a nearly homologous hemolysin determinant (pANN202-312) in a minicell-producing background, we identified where along the hemolysin determinant the different polypeptide species were encoded. These results are shown in Fig. 3. The order, size, and location of the 19.7- and 109.9-kd species corresponded, respectively, to the *hlyC* and *hlyA* cistrons (9). Although we confirmed the apparent existence of two genes downstream of *hlyA* due to the presence of two large ORFs, the order and size of the 79.9- and 54.6-kd polypeptides were not similar to those in previous reports (13, 18, 42). Because of this complication and the desire to utilize conventional genetic nomenclature, we propose that the two cistrons, respectively, be termed *hlyB* and *hlyD* (rather than *hlyBa* and *hlyBb*).

The minicell analysis of Tn*1* insertions within the *hlyB* region of pSF4000 did not reveal the loss of any polypeptide species detected visually in autoradiographs (data not shown). Therefore, we chose to examine the polypeptides

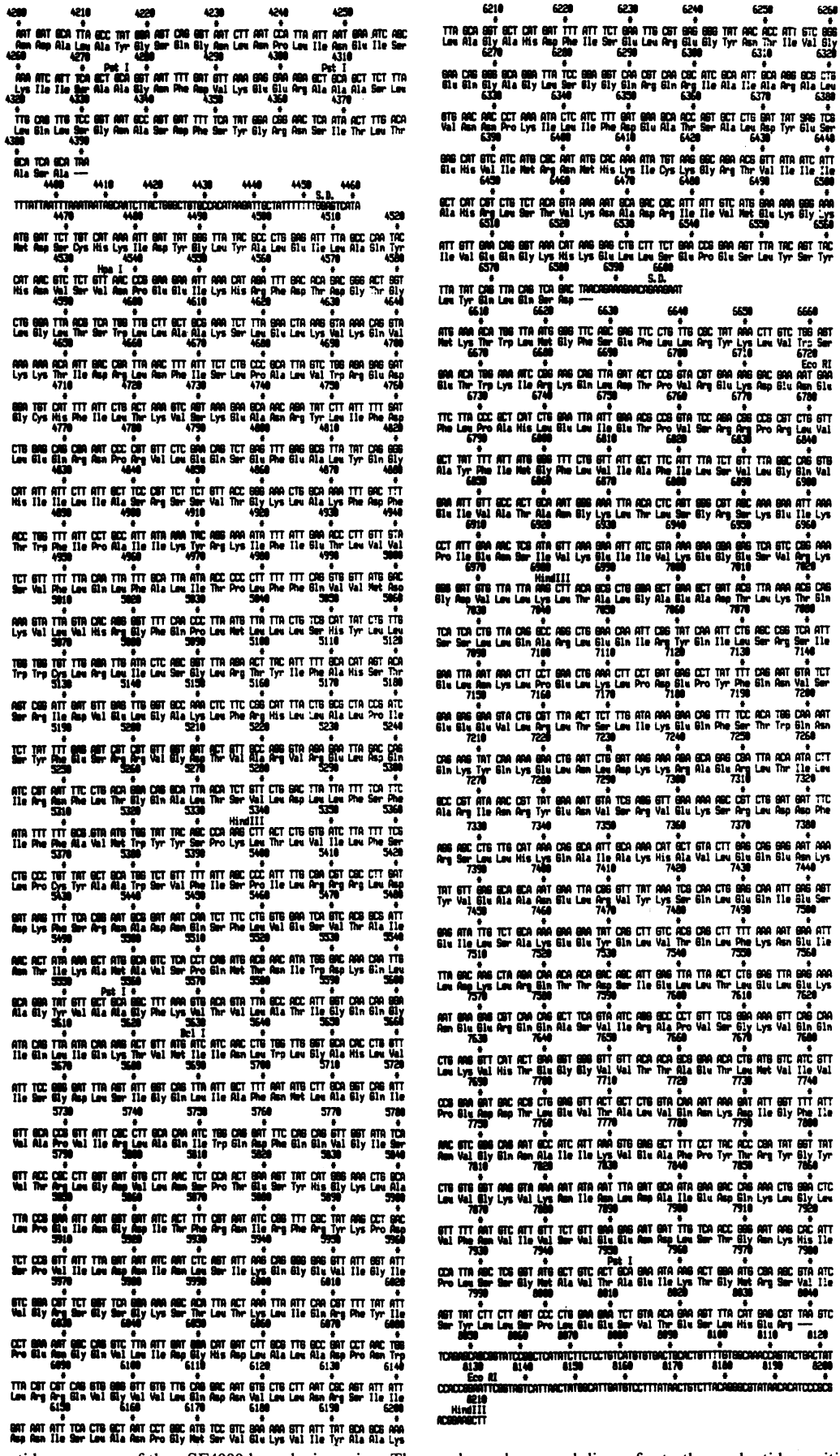
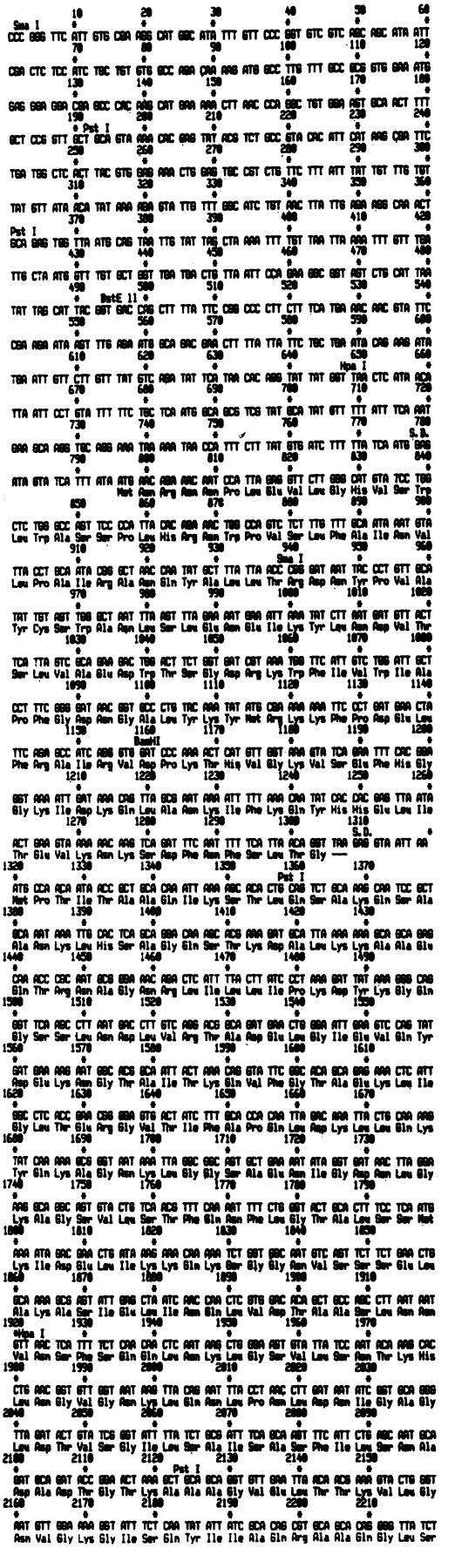


FIG. 2. Nucleotide sequence of the pSF4000 hemolysin region. The numbers above each line refer to the nucleotide position. Also listed above the nucleotide sequence are selected restriction endonuclease cleavage sites that are referred to in the text. The predicted amino acid sequences for HlyC (bp 796 to 1,303), HlyA (bp 1,320 to 4,388), HlyB (bp 4,462 to 6,582), and HlyD (bp 6,604 to 8,037) are shown beneath the DNA sequence. The positions for possible ribosome-binding sequences (Shine-Dalgarno [S.D.] sequences of the mRNA) are shown above the DNA sequence preceding each gene.


Sma I
Pst I
BstE II
Hpa I
S.D.
Pst I
BamHI
Sma I
Hpa I
Eco RI
Hpa II
Bgl II

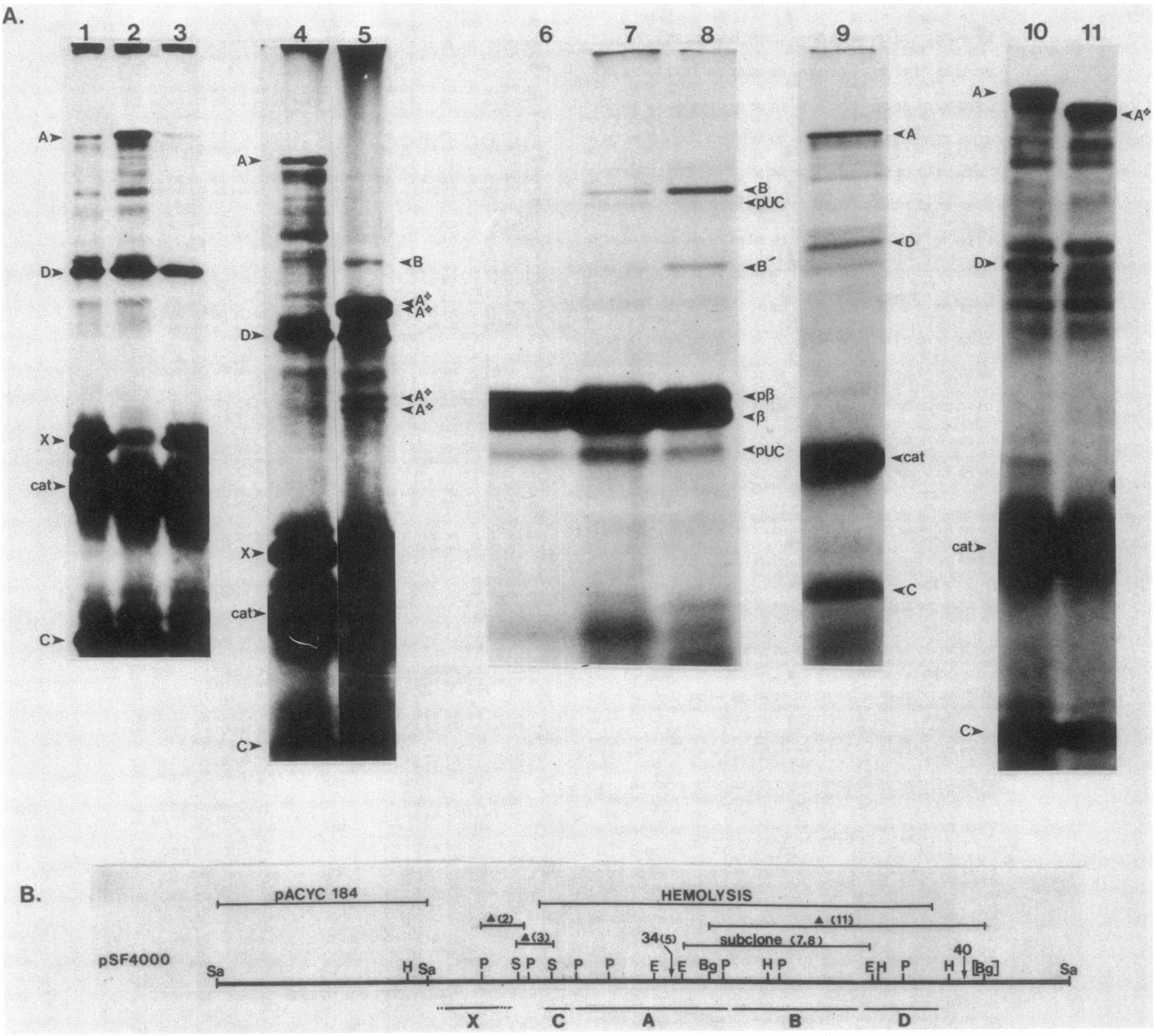


FIG. 3. Polypeptides encoded by different plasmids in minicells. (A) Fluorograms of sodium dodecyl sulfate-10% polyacrylamide gels containing [^{35}S]methionine-labeled polypeptides encoded by different hemolysin recombinant plasmid derivatives present in purified *E. coli* minicells. (B) Physical locations of deletions (▲), Tn*1* insertions (↓), and subclones of hemolysin recombinant plasmids. The numbers enclosed within parentheses indicate the lane in (A) which contains the polypeptides encoded by that particular plasmid derivative. In (A) lanes 1 and 4 show polypeptides encoded by pSF4000. The lettered arrows next to these lanes indicate the location of the 110-kd HlyA (A), 54-kd HlyD (D), 33- and 32-kd X polypeptides (X), chloramphenicol acetyl transferase (cat), and the 19-kd HlyC polypeptide (C). Lanes 2 and 3 show the pSF4000Δ*Pst*I and pSF4000Δ*Sma*I deletion derivatives. Lane 5 contains the plasmid pSF4000::Tn*1*(34). The A* bands represent the truncated form of the 110-kd HlyA protein and its probable proteolytic breakdown products. B indicates the location of the HlyB polypeptide. Lane 6 contains pUC9. Lanes 7 and 8 are pUC9-based *Eco*RI subclones of pANN202-312 in which both orientations of the insert (pWAM326 and pWAM327) have been examined. Polypeptides (B and B') of 77 and 46 kd are specifically encoded by these plasmids. pβ and β designate the precursor and mature forms of the pUC9 β-lactamase, and the pUC arrows represent pUC9-associated polypeptides. Lanes 9 and 10 contain pANN202-312. Lane 11 shows the plasmid pANN202-312Δ*Bgl*II (pWAF185). In the restriction map [Bg] represents a *Bgl*II site just outside the hemolysin determinant that is present in pANN202-312 but not pSF4000. Again, the A* represents a truncated HlyA product. The horizontal bars below the restriction endonuclease fragment map show the assignment of the encoding regions for HlyC, HlyA, HlyB, and HlyD. X designates the two polypeptides outside of the hemolysin region whose function is unknown, but which probably represent a precursor-product relationship (R. A. Welch, unpublished data).

synthesized by subclones of the *hlyB* region. It had previously been reported that an *Eco*RI fragment spanning the *hlyB* region of pANN202-312 inserted into pUR222 (pANN250-222) encoded a 46-kd polypeptide in minicells (13). We constructed similar plasmids (pWAM326 and pWAM327) and found that not only did they encode a 46-kd polypeptide but a 77-kd polypeptide as well (Fig. 3, lanes 7 and 8). The orientation of the *Eco*RI fragment in pUC9

(pWAM326 vs pWAM327) did not appear to influence the amount of either polypeptide relative to the other. It is apparent that isopropyl-β-D-thiogalactoside induction of the *lac* promoter present upstream of the *Eco*RI insertion site in pWAM327 enhanced equally the synthesis of the 77- and 46-kd polypeptides (Fig. 3, lane 8). The DNA sequence analysis of the *Eco*RI fragment indicated that there would not be a translational fusion created with the pUC9 *lac* sequences at the *Eco*RI insert for either the *hlyA* or *hlyD* gene. This was further substantiated by the fact that for both orientations of the *Eco*RI fragment in pUC9, all of the detectable polypeptides remained similar in size. Although our sequence data was derived from pSF4000, DNA sequence comparison of pSF4000 and pANN202-312 (*hlyC* and the first 120 bp of *hlyA*) revealed a 97% homology (R. A. Welch and S. Pellett, manuscript in preparation). Additional evidence for synthesis of the 79.9-kd HlyB polypeptide came from examination of the minicell results with the plasmid pSF4000::Tn*1*(34) (Fig. 3, lane 5). A polypeptide ca. 77 kd in size was observed when the *hlyA* reading frame was interrupted by the transposon insertion. Therefore, because a 79.9-kd polypeptide is predicted to be encoded by the pSF4000 *hlyB* region and a polypeptide close in size is encoded by the *hlyB* region of pANN202-312, we think that a correct assignment has been made.

The inability to detect the loss of the HlyB polypeptides with transposon insertions within *hlyB* appears to be due to two factors. First, there apparently was low production of HlyB compared with HlyC, HlyA, and HlyD (Fig. 3, lane 4 versus lane 5 and lane 10 versus lane 11). Second, there was probable masking of the HlyB species by what are thought to be proteolytic breakdown products of HlyA that are abundant in the minicell background (6).

We employed a derivative of pANN202-312 to assign the 54-kd polypeptide to the *hlyD* region. pSF4000 and pANN202-312 shared a *Bgl*II site (bp 3,807; Fig. 2). pANN202-312 had an additional *Bgl*II site ~5 kb in the downstream direction, which lay ca. 600 bp distal to the C terminus of HlyD (44). A deletion between the two *Bgl*II sites was isolated in vitro, and this plasmid (pWAF185) was transformed into the DS410 minicell background. In lane 11 of Fig. 3 we see that this resulted in the loss of HlyA and the appearance of a probable truncated form of HlyA. In addition, a 52- to 54-kd polypeptide species was missing. This species is felt to represent HlyD because a similar polypeptide, based on comigration, was seen for pSF4000. The location of a Tn*1* insertion near the *hlyD* region of pSF4000, which does not interrupt hemolysin production, is indicated in Fig. 3B by the arrow numbered 40. When this plasmid was examined in minicells, the 54-kd polypeptide was still evident (data not shown). Based on the DNA sequence analysis and the minicell results with the *Bgl*II deletion plasmid and the Tn*1*(40) insertion, it is likely the assignment of the 54-kd polypeptide to the region shown on pSF4000 physical map is correct. The identification of the 54-kd HlyD polypeptide and its transcriptional order relative to the other cistrons is in disagreement with findings made in the laboratory of Goebel (13, 18). Our results concerning HlyD have recently been confirmed by N. Mackman and B. Holland (personal communication).

DNA sequence features. The first general observation to be made was the relatively low guanine-plus-cytosine content (40.2%) of this DNA sequence when compared with that for the *E. coli* genome (50 to 52%) (25). This situation is similar to that seen for the *E. coli* heat-labile enterotoxin A and B subunits (46). The lower guanine-plus-cytosine content in that instance was associated with a codon usage pattern distinct from that seen for a number of sequenced *E. coli* genes. Shown in Fig. 4 is a codon preference plot of the four hemolysin genes based on the codon frequencies observed for a large collection of sequenced genes from *E. coli* (10). From these DNA sequences, Gribskov et al. derived an optimal codon usage pattern for *E. coli*. This permits the derivation of a codon preference statistic for each position in each of the three reading frames. A sliding window along each reading frame of 25 codons permits a statistical measure of the similarity in occurrence of that set of codons to that predicted from the optimal codon usage table. The plots for the four hemolysin genes revealed that there was a relatively poor match in the hemolysin codon usage with the optimal codon usage for *E. coli* genes. For both highly and moderately expressed *E. coli* genes, the codon preference plot is often well above the random codon preference statistic (10). Also shown is the occurrence of codons that occur 5% or less of the time for *E. coli* in each of the synonymous codon families (Fig. 4). In *E. coli* these rare codons occur frequently in the nonencoding regions of DNA sequences. For the four hemolysin genes, the rare codons occurred as frequently in the ORFs as elsewhere in the sequence.

Presented in Fig. 5 are the results of a deletion analysis of the region upstream of the initial cistron, *hlyC*. The 132-bp region between the *Pst*I and *Bst*EII sites was necessary for hemolysin synthesis. There was no sizable ORF extending across the *Bst*EII site through to the putative translational start of *hlyC*. Although not shown, Tn*1* insertions to the left of the *Pst*I site did not have any detectable effect on the hemolysin phenotype, whereas those just to the right of this site caused a reduction in the hemolysis zone size (45). In fact, cells harboring the *Bst*EII deletion did show a slight zone of hemolysis surrounding colonies after 48 h of incubation of blood agar plates. In the lower portion of Fig. 5 is shown the 437-bp region beginning at the *Pst*I site (bp 363) through to the putative translational start of HlyC (bp 6,796). Indicated within this DNA sequence are a number of features of possible consequence. There are numerous relatively small inverted repeats (3 to 7 bp), as well as several directly repeated sequences clustered between the *Pst*I and *Bst*EII sites.

A search within the DNA sequence from the PstI site (bp 363) to the beginning of *hlyC* for subsequences similar to those described as consensus assignments for *E. coli* promoters showed that there are in fact numerous potential transcriptional initiation sites (14, 34). Identified in Fig. 5 are possible promoter sequences based on this search. The subsequences shown are only those meeting the following criteria: seven or more matches with the 12 nucleotides making up the −10 (T*A*TAAT*) and −35 (TT*G*ACA) hexanucleotides (with perfect matches always at the nucleotides marked with an asterisk), a −35 to −10 nucleotide spacing of 17 ± 2 bp, and a purine residue 5 to 7 bp downstream from the last T of the −10 hexanucleotide (representing the putative intitial nucleotide of the transcript). Based on these relatively stringent conditions, we made six different promoter-like sequence assignments upstream of *hlyC*. Two of these (− 10, bp 407; −10, bp 468) reside within the *Pst*I-to-*Bst*EII region.

The fact that all four ORFs were encoded sequentially on one DNA strand suggests that all four genes could be translated from one large transcript initiated upstream of *hlyC*. However, some DNA sequence features and minicell results suggest a more complicated operon configuration. In

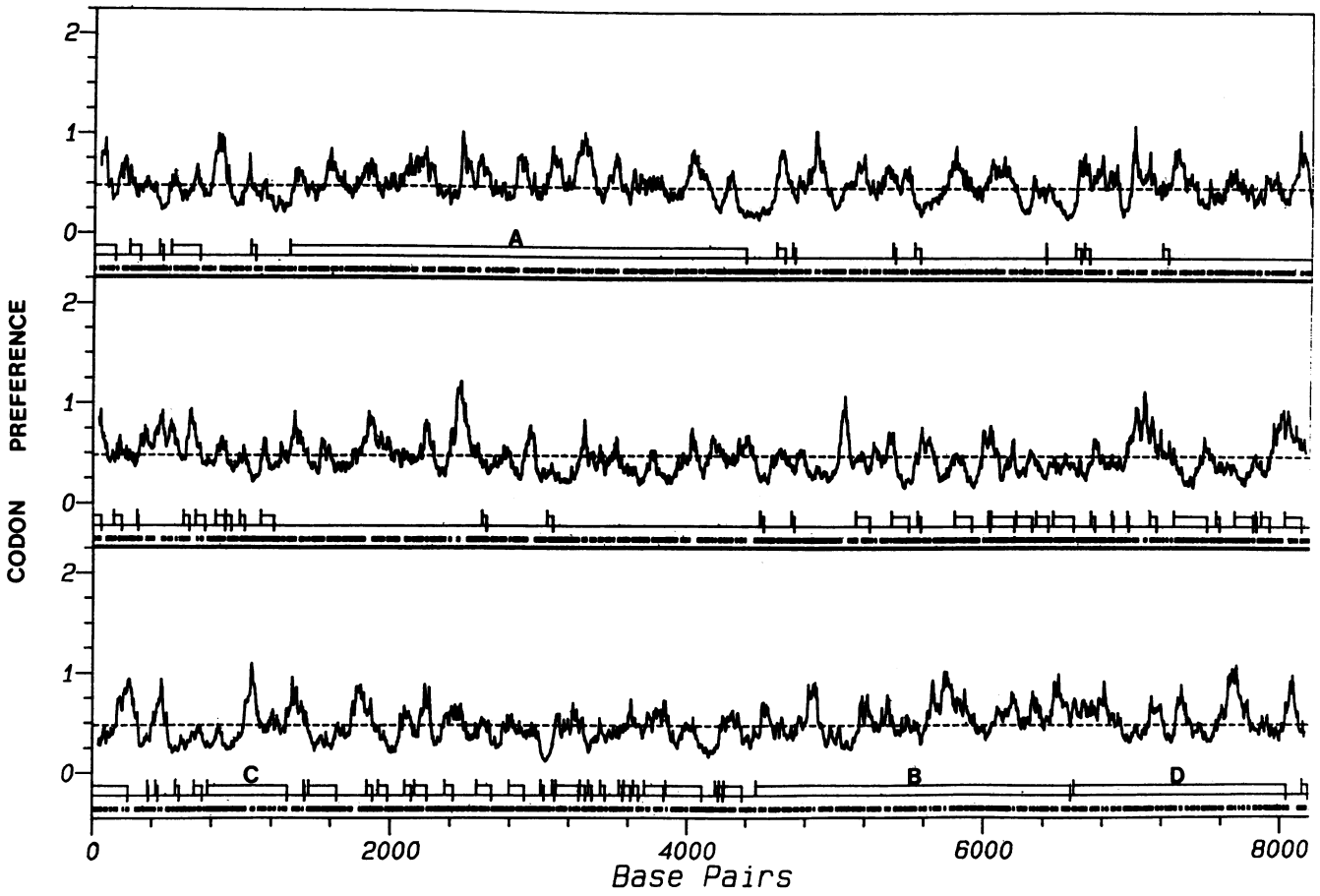


FIG. 4. Codon preference plot of the *E. coli* hemolysin. The plot is divided into three panels, with each representing a different reading frame. The horizontal dashed line shows the calculated codon preference statistic for the theoretical random sequence of the same base composition as the codon frequency table. The four ORFs (C, A, B, and D) are shown as labeled open boxes beneath the plotted codon preference statistic. Other ORFs are marked by vertical lines not crossing the horizontal line (AUG codons), followed by an open box, and the stop codons are shown as vertical lines crossing the horizontal line. The short vertical bars beneath the ORFs indicate the occurrence of rare *E. coli* codons (5% or less within a synonomous family).

Fig. 6 we show a possible mRNA structure predicted from the DNA sequence just downstream of the apparent C terminus of *hlyA*. This structure is very similar in architecture to *rho*-independent transcriptional termination regions identified for a number of *E. coli* genes (34). Often these structures are depicted as long-base-paired stems, followed by four to eight unpaired uridylates. We show in Fig. 6 a 21-bp stem that could be formed if base pairing with the uridylates occurs. Because AU pairs do not contribute as much stability to a duplex as GC pairs (39), a shorter 16-bp stem followed by seven unpaired uridylates, still would have a favorable free energy of formation (23.6 kcal [ca. 98.7 kJ]/mol). In conjunction with this observation, we also note that the amount of the HlyB polypeptide relative to HlyC and HlyA polypeptides was lower based on the intensity of the autoradiograph signal in gels of minicells (Fig. 3, lane 4 versus lane 5). The low signal intensity cannot be explained on the basis of relative low methionine content because there were more methionines predicted in HlyB than in HlyA (lane 12 versus lane 5).

An examination of the DNA sequence surrounding the possible terminator region did reveal a close match to the consensus sequences for *E. coli* promoters. This region is indicated in Fig. 6B. In the absence of induction of the *lac* promoter, the orientation of the *Eco*RI fragment cloned in pUC9 did not appear to influence the relative amount of expression of the B and B′ polypeptides based on the intensity of the autoradiographic signal (data not shown). This is evidence suggesting that *hlyB* has its own promoter.

Goebel and co-workers subcloned the 5-kb *Bgl*II pANN202-312 fragment (pANN205-222) and demonstrated that the orientation of the fragment did not affect the ability to complement successfully a Tn*5* insertion mutation in the most distal transport function (42). This suggested that the distal hemolysin transport-associated function is under the control of its own promoter (42). We observed a more abundant expression of HlyD relative to HlyB (Fig. 3, lanes 4 and 5). Under these circumstances, HlyD expression was clearly not polar to hlyB and confirmed the strong likelihood of an HlyD-specific promoter. Three possible matches to the consensus sequence for *E. coli* promoters based on the criteria we have chosen exist for *hlyD* within the *hlyB*-encoding sequence. The −10 regions for these sequences reside at bp 4,589, 4,824 and 5,211 (see Fig. 2).

Predicted amino acid sequence features. Based on the predicted amino acid content, HlyC and HlyB would be basic proteins (isoelectric points of 9.5 and 10.2), whereas HlyA and HlyD would be slightly acidic (isoelectric points of 6.1 and 6.3). There are more charged amino acids per unit length for HlyD (1 per 3.3 amino acids) than for HlyC (1 per

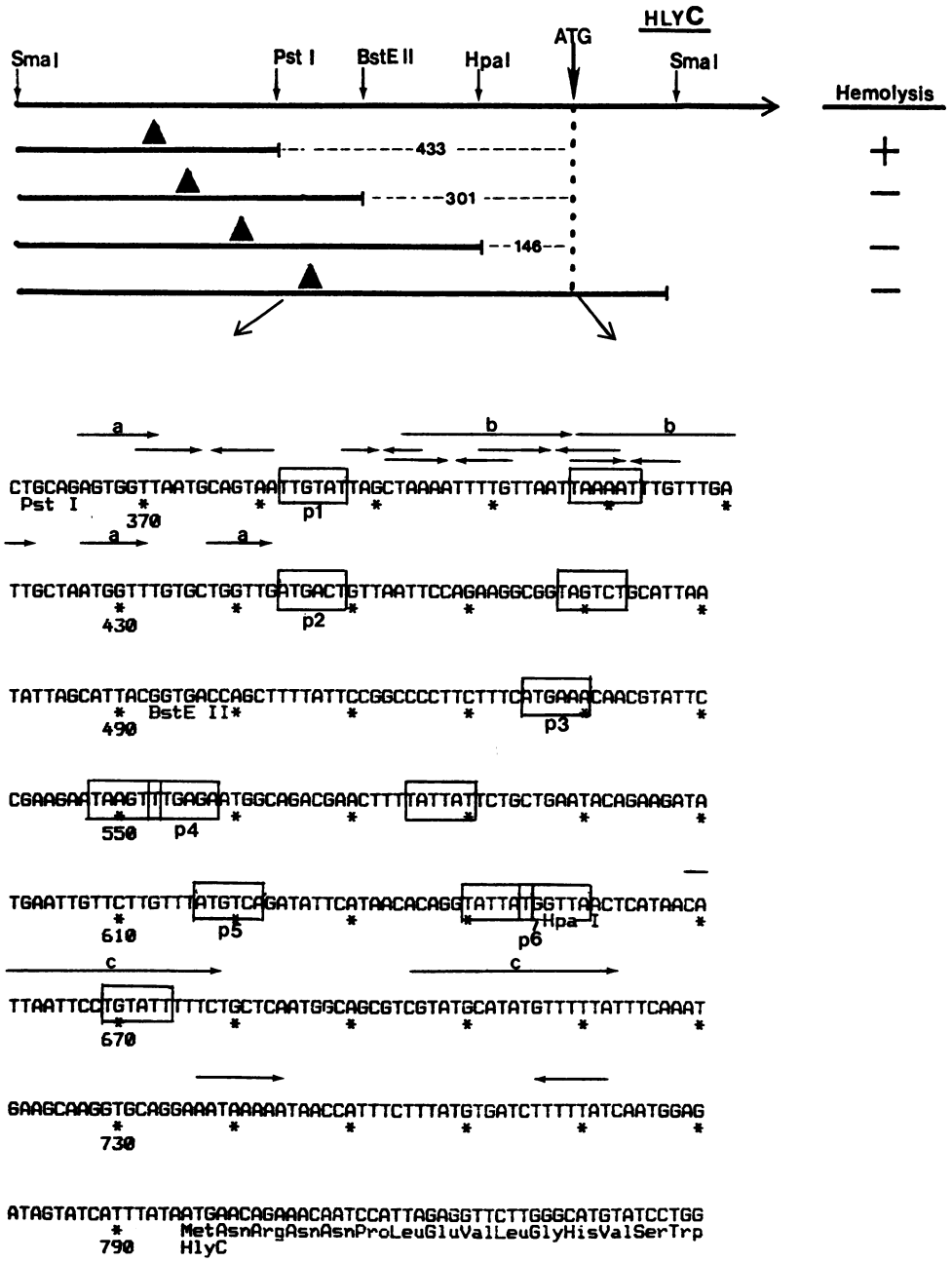


FIG. 5. Deletion and DNA sequence analysis of the region upstream of HlyC. At the top is a restriction endonuclease fragment map of the region upstream of HlyC. The four horizontal bars beneath the map indicate the rightward extent of in vitro-derived pSF4000 deletions. The numbered dashed lines indicate the number of bp from the rightward endpoint of the deletion to the initial HlyC ATG codon. The plus and minus designations beneath "Hemolysis" indicate whether these particular deletion derivatives still encoded a hemolytic phenotype on blood agar plates. The DNA sequence listed below the deletion map begins with the *PstI* site (Fig. 2; bp 363) and ends 45 bp after the beginning of the HlyC sequence. The first 15 predicted amino acids of HlyC are listed below the DNA sequence. The numbered asterisks refer to the bp number of this sequence as listed in Fig. 2. The lettered horizontal arrows indicate direct DNA sequence repeats, whereas the unlettered arrows pointing at one another indicate inverted repeat sequences. The boxes labeled p1, p2, etc., designate hexamers very similar in sequence to the consensus sequence for the −35 region of *E. coli* promoters. The unlabeled box 17 ± 2 bp downstream from the −35 hexamers represents an additional hexamer very similar to the −10 consensus sequence for *E. coli* promoters.

3.8 amino acids), HlyA (1 per 3.8 amino acids), and HlyB (1 per 4.3 amino acids). A curious finding involving HlyA was the regional feature of basic versus acidic amino acids in its primary sequence. Amino acids 1 to 230, 231 to 430, and 431 to 1,023 had estimated isoelectric points of 10.2, 6.8, and 5.6, respectively. An interesting feature of the predicted amino acid content was the absence of cysteine residues in HlyA and HlyD. Otherwise, we did not find an abundance or lack of any particular amino acid or class of amino acids.

Shown in Fig. 6 are the successive predicted hydropathy plots for the sequence of amino acids of each hemolysin gene based on the prediction of Kyte and Doolittle (20). Super-

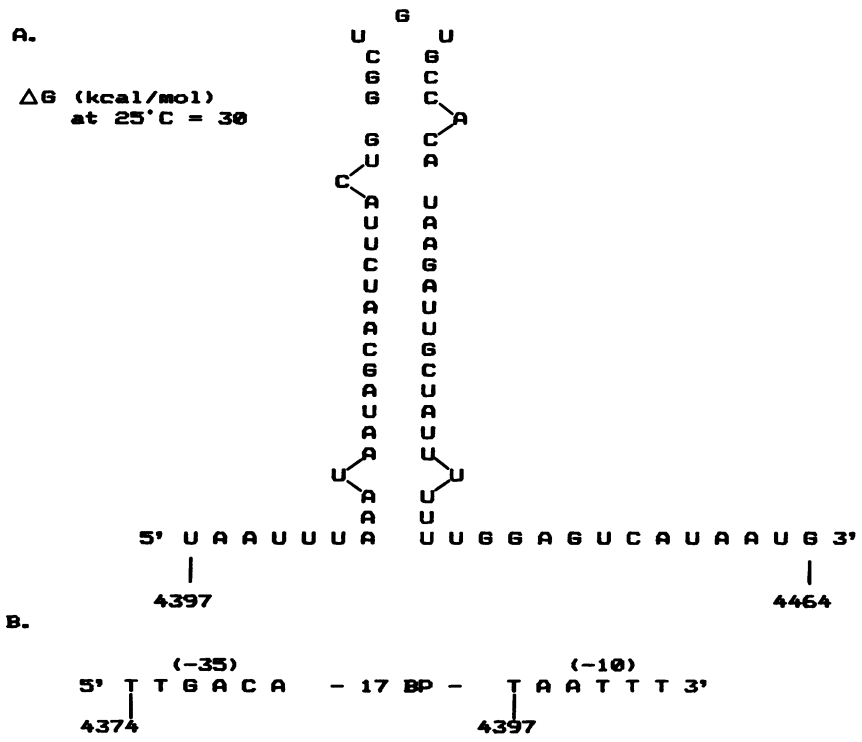


FIG. 6. Predicted mRNA structure and promoter-like DNA sequence distal to the HlyA C terminus. (A) Predicted secondary structure in the mRNA downstream of *hlyA*. The numbers beneath the sequence indicate the numbers of the corresponding DNA bp as presented in Fig. 2. The 21-bp stem and its calculated free energy of formation are depicted. Shown is the longest predicted stem structure within this region where it would incorporate five of the seven consecutive uridylates. The 3' end of the mRNA sequence shown represents the tentative HlyB initiation codon. (B) Promoter-like DNA sequence immediately upstream of the stem-loop structure.

imposed on each hydropathy plot is the predicted amino acid secondary structure based on the rules of Chou and Fasman (5). An examination of the predicted N-terminal portion of each protein revealed the absence of a pattern similar to that associated with the signal sequence for secreted proteins (27, 31, 37, 41). Both positively and negatively charged residues were present, and there was no core of strongly hydrophobic amino acids in the predicted sequences.

The hydropathy plots are useful in identifying potential transmembrane hydrophobic domains, as well as potential antigenic sites (20). Large uncharged hydrophobic regions (≥15 amino acids) were apparent in HlyC (residues 25 to 40), HlyA (residues 140 to 160, 245 to 260, 295 to 325, and 375 to 405), HlyB (residues 155 to 180), and HlyD (residues 55 to 80). Charged hydrophilic regions representing potential antigenic sites existed at the HlyC C terminus (positively charged), the N terminus of HlyA (positively charged), and the entire C-terminal third of HlyA, HlyB (residues 315 to 330 and the C terminus), and HlyD (residues 15 to 40, 190 to 215, and 318 to 330).

DISCUSSION

We present the DNA sequence for an *E. coli* hemolysin which was molecularly cloned from the chromosome of an O4 serotype, uropathogenic isolate of *E. coli*. One strand of the DNA possessed four successive ORFs which physically coincide with four reported hemolysin cistrons (29, 30, 42). We also present evidence for the existence of pSF4000-encoded polypeptides similar in size to each of the predicted proteins. The apparent size of the HlyC and HlyA polypeptides coincides with previous reports (9). However, the evidence we present for the molecular mass of the two hemolysin transport genes does not coincide. It has been reported the HlyB (or HlyBa) and HlyD (or HlyBb) proteins are 46 and 64 kd in molecular mass (12). Albeit that our DNA sequence was derived from a hemolysin determinant of different origin than that studied in the laboratory of Goebel, our analysis of pANN202-312-encoded polypeptides supports the pSF4000 DNA sequence data that two proteins of 79.9 and 54 kd are encoded by the *hlyB* and *hlyD* regions, respectively.

It is curious that in minicells we detected both 77- and 46-kd polypeptides for the *Eco*RI fragment covering the *hlyB* cistron. Neither protein can be the result of translational fusions between *hlyA* or *hlyD* sequences and the *lac* region bordering the *Eco*RI site of pUC9. There were no sizable ORFs in any of the other five possible reading frames of the *Eco*RI fragment. Thus, we are left with two possible explanations. The 46-kd B' polypeptide (Fig. 3, lane 8) may be a proteolytic cleavage product of polypeptide B. Alternatively, there may be different translational starts within the ORF present in this fragment. We noted that at bp 5,317 there was an ATG codon preceded by a potential ribosome-binding site (GCGG). The predicted molecular mass of a protein covering the ORF from bp 5,317 to 6,582 is 46 kd, and the sequence of the first 21 amino acids of that potential product has a signal-like sequence (27, 31, 37, 41). At this time, we do not know which alternative explanation is true. Also, we do not know whether the hemolysin transport function assigned to the *HlyB* cistron region is carried out by either or both of the 77- and 46-kd proteins.

Although we do not provide direct evidence for the existence of different hemolysin-specific mRNA species, a number of DNA sequence features and minicell results allow

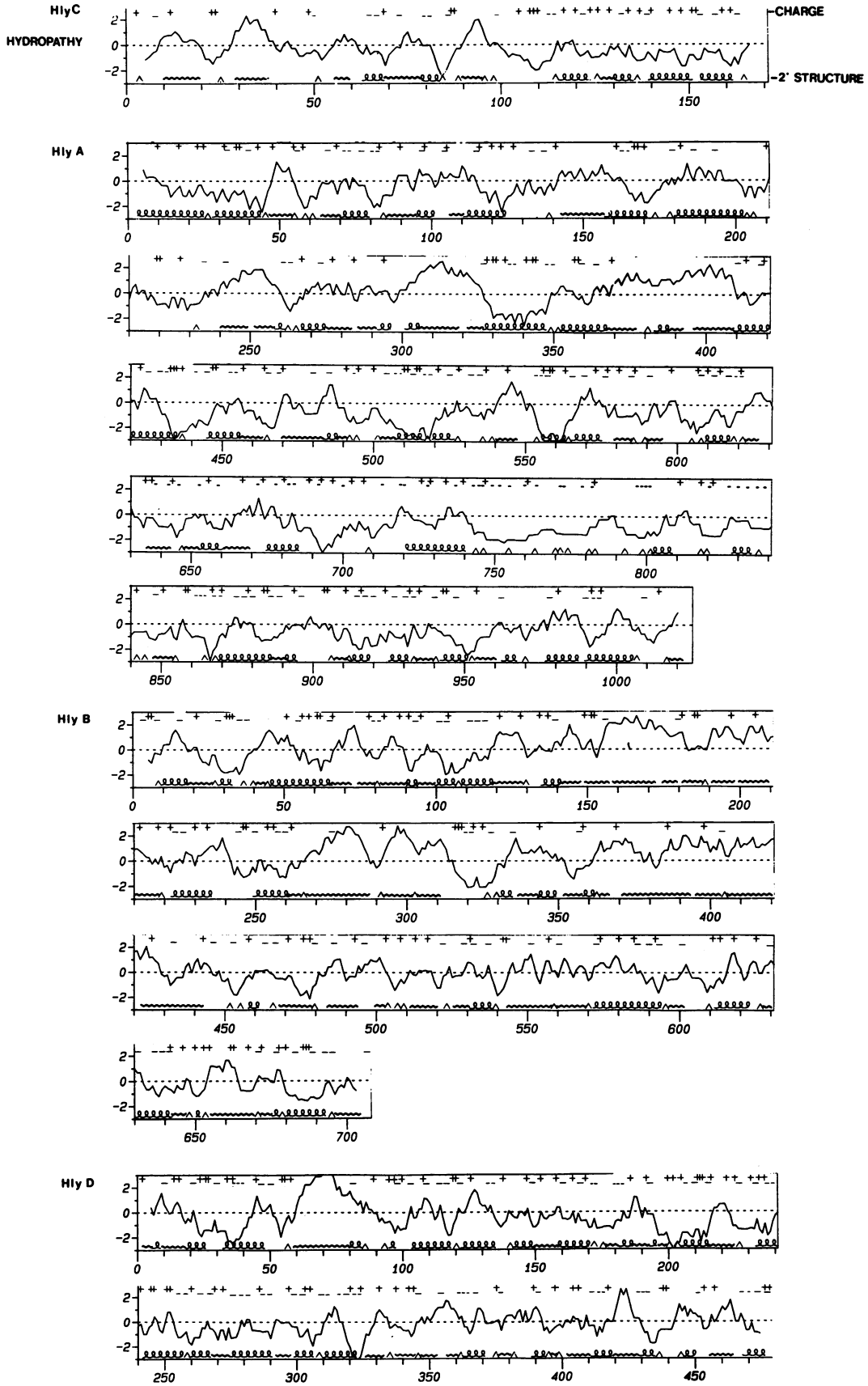


FIG. 7. Predicted hydropathy profile and secondary structure of the four hemolysin genes. The four sets of panels represent the combined hydropathy and secondary structure predictions of the deduced amino acid sequences for HlyC, HlyA, HlyB, and HlyD. The numbered horizontal axis under each panel represents the amino acid number. The left vertical axis indicates the relative hydrophobicity (positive ordinate) or hydrophilicity (negative ordinate). Plotted is the calculated hydropathy value for a window of nine amino acids as the frame moves consecutively one amino acid at a time toward the C terminus (20). Listed horizontally along the top of each panel are plus or minus signs indicating the occurrence of acidic and basic amino acids, respectively. Shown just above the bottom horizontal axis are the predicted secondary structures calculated according to Chou and Fasman (5). The structures are alpha helix (⌇⌇⌇), β sheet (〰), and turns (/\). The undesignated gaps are presumed to be random coils.

for some predictions about the transcriptional organization of the hemolysin determinants. It is possible that a single polycistronic mRNA species may encode all four genes. This seems unlikely because we observed in minicells higher amounts of the HlyD protein relative to the HlyB species. The identification of a potential mRNA secondary structure between the *hlyA* and *hlyB* genes that is similar to structures encoded in regions of known transcriptional termination suggests that occasional readthrough into the *hlyB* region may account for the lower expression of *hlyB* relative to *hlyA*. Alternatively, the *hlyA* transcript may efficiently end at this site, and a second *hlyB*-specific transcript may begin from a weaker promoter. The ability to detect the 77- and 46-kd polypeptides in either orientation of the *Eco*RI fragment in subclones covering the *hlyB* region suggests that a promoter specific for *hlyB* may exist. The promoter-like sequence shown in Fig. 6B is presented only as a potential candidate.

The HlyD protein is present in quantities greater than that of the HlyB protein. This can best be accounted for by the presence of a strong promoter within the *hlyB*-encoding sequence. There is insufficient space in the *hlyB* and *hlyD* intercistronic region for the needed transcription initiation signals. The existence of a *hlyD*-specific promoter has been supported by complementation data (42).

The in vitro deletions of the area upstream of *hlyC* indicated that a region more than 300 bp from the putative HlyC start codon was required for hemolysin synthesis. Our analysis of the DNA sequence of this region indicated that no ORFs existed in this area. The presence of numerous direct and indirect repeated DNA sequences suggested a regulatory site. The significance of any of these sequences is untested and awaits the isolation of additional mutants in this region. We present six different subsequences as possible candidates for *hlyC* promoters. Because we cannot identify a terminator-like sequence between the *hlyC* and *hlyA* genes, we assume that a single mRNA encodes HlyC and HlyA. Overall, the tentative model we propose for the transcriptional organization of the hemolysin determinant dictates that there are three transcriptional units (*hlyC-hlyA*, *hlyB*, and *hlyD*). Work in our laboratory is presently directed at the identification of the in vivo transcriptional start sites for each of the predicted transcripts.

DNA-DNA hybridization experiments have indicated that the *E. coli* hemolysin determinant is unique to a limited number of *E. coli* isolates (45). In an evolutionary sense, we suggested that this was evidence that the *E. coli* hemolysin only recently came to reside in *E. coli*. We provide further evidence for this conjecture based on the discrepancy in guanine-plus-cytosine content between the *E. coli* genome and the hemolysin sequence. In addition, the codon usage pattern for the *E. coli* hemolysin is unlike that of other *E. coli* genes (10, 11, 16). In the hemolysin-encoding genes there is frequent use of rare *E. coli* codons. Therefore, it seems likely that the hemolysin came from an organism not closely related to *E. coli*.

In an accompanying paper, we provide evidence that the *E. coli* hemolysin is secreted extracellularly (6). Analysis of the amino acid content of extracellular proteins from a variety of gram-positive and gram-negative genera led Pollock and Richmond (32) to note that there are few if any cysteine residues in this class of proteins. It is interesting that the predicted absence of cysteine in the hemolysin structural protein conforms to their observation.

The N-terminal amino acid sequence is known only for the hemolysin structural gene (6). The other hemolysin proteins (HlyC, HlyB [B'], and HlyD) have not been sufficiently purified to permit this analysis. There is little physical evidence concerning their location in the cell, although it has been suggested the HlyC protein is in the cytoplasm and HlyB and HlyD proteins are in the outer membrane (42). The significance of the observation that signal-like sequences appear to be absent for HlyB and HlyD is unknown and awaits their purification and localization.

We note that based on the hydropathy plots, potential membrane spanning domains (20) occur within each protein. We are using a variety of genetic and biochemical approaches to localize each protein in the cell. In turn, we intend to delineate a structural and functional role for each hemolysin gene in the apparent hemolysin secretory pathway. In addition, we predict that there may be a complex set of interactions between the hemolysin protein and different eucaryotic cells. We are examining the possible existence of discrete domains within the hemolysin protein that may be involved in this process.

ACKNOWLEDGMENTS

We thank Mike Gribskov and Dick Burgess for help in use of the protein sequence programs. In addition, we thank Bernie Weisblum for his visits to our laboratory and Caroline Fritsch and Christine Crawford for their help in the preparation of this manuscript.

This work was supported by Public Health Service grant AI-20323 from the National Institutes of Health. Additional support came from the University of Wisconsin Graduate School, the University of Wisconsin Medical School, and the Shaw Fund. T.F., a predoctoral student in the Molecular Biology Program, was supported by funds from the University of Wisconsin Graduate School.

LITERATURE CITED

1. **Berger, H., J. Hacker, A. Juarez, C. Hughes, and W. Goebel.** 1982. Cloning of the chromosomal determinants encoding hemolysin production and mannose-resistant hemagglutination in *Escherichia coli*. J. Bacteriol. **152**:1241–1247.
2. **Cavalieri, S. J., and I. S. Snyder.** 1982. Effect of *Escherichia coli* alpha-hemolysin on human peripheral leukocyte viability in vitro. Infect. Immun. **36**:455–461.
3. **Cavalieri, S. J., and I. S. Snyder.** 1982. Effect of *Escherichia coli* alpha-hemolysin on human peripheral leukocyte function in vitro. Infect. Immun. **37**:966–974.
4. **Cavalieri, S. J., and I. S. Snyder.** 1982. Cytotoxic activity of partially purified *Escherichia coli* alpha hemolysin. J. Med. Microbiol. **15**:11–21.
5. **Chou, P. Y., and G. D. Fasman.** 1974. Prediction of protein conformation. Biochemistry **13**:222–245.
6. **Felmlee, T., S. Pellett, E.-Y. Lee, and R. A. Welch.** 1985. *Escherichia coli* hemolysin is released extracellularly without cleavage of a signal peptide. J. Bacteriol. **163**:88–93.
7. **Gadeberg, O., and I. Orskov.** 1984. *In vitro* cytotoxic effect of α-hemolytic *Escherichia coli* on human blood granulocytes. Infect. Immun. **45**:255–260.
8. **Gill, R. E., F. Heffron, and S. Falkow.** 1979. Identification of the protein encoded by the transposable element Tn*3* which is required for its transposition. Nature (London) **282**:797–801.
9. **Goebel, W., and J. Hedgpeth.** 1982. Cloning and functional characterization of the plasmid-encoded hemolysin determinant of *Escherichia coli*. J. Bacteriol. **151**:1290–1298.
10. **Gribskov, M., J. Devereux, and R. R. Burgess.** 1984. The codon preference plot: graphic analysis of protein coding sequences and prediction of gene expression. Nucleic Acids Res. **12**: 539–549.
11. **Grosjean, H., and W. Fiers.** 1982. Preferential codon usage in procaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. Gene **18**:199–209.
12. **Hacker, J., C. Hughes, H. Hof, and W. Goebel.** 1983. Cloned hemolysin genes from *Escherichia coli* that cause urinary tract

infection determine different levels of toxicity in mice. Infect. Immun. **42**:57–63.
13. **Hartlein, M., S. Schiessl, W. Wagner, U. Rdest, J. Kreft, and W. Goebel.** 1983. Transport of hemolysin by *Escherichia coli*. J. Cell. Biochem. **22**:87–97.
14. **Hawley, D. K., and W. R. McClure.** 1983. Compilation and analysis of *Escherichia coli* promoter DNA sequences. Nucleic Acids Res. **11**:2237–2255.
15. **Hull, S. I., R. A. Hull, B. H. Minshew, and S. Falkow.** 1982. Genetics of hemolysin of *Escherichia coli*. J. Bacteriol. **151**:1006–1012.
16. **Ikemura, T.** 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *Escherichia coli* translational system. J. Mol. Biol. **151**:389–409.
17. **Jaurez, A., and W. Goebel.** 1984. Chromosomal mutation that affects excretion of hemolysin in *Escherichia coli*. J. Bacteriol. **159**:1083–1085.
18. **Jaurez, A., C. Hughes, M. Vogel, and W. Goebel.** 1984. Expression and regulation of the plasmid-encoded hemolysin determinant of *Escherichia coli*. Mol. Gen. Genet. **197**:196–203.
19. **Knapp, S., J. Hacker, I. Then, D. Müller, and W. Goebel.** 1984. Multiple copies of hemolysin genes and associated sequences in the chromosomes of uropathogenic *Escherichia coli* strains. J. Bacteriol. **159**:1027–1033.
20. **Kyte, J., and R. F. Doolittle.** 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. **157**:105–132.
21. **Laemmli, U. K.** 1970. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature (London) **227**:680–685.
22. **Levy, S. B.** 1974. R factor proteins synthesized in *Escherichia coli* minicells: incorporation studies with different R factors and detection of deoxyribonucleic acid-binding proteins. J. Bacteriol. **120**:1451–1463.
23. **Mackman, N., and B. Holland.** 1984. Secretion of 107-k dalton polypeptide into medium from a haemolytic *Escherichia coli* K12 strain. Mol. Gen. Genet. **193**:312–315.
24. **Maniatis, T., E. F. Fritsch, and J. Sambrook.** 1982. Molecular cloning. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
25. **Marmur, J., and P. Doty.** 1962. Determination of the base composition of deoxyribonucleic acid from its thermal denaturation point. J. Mol. Biol. **3**:585–594.
26. **Messing, J., R. Crea, and P. H. Seeburg.** 1981. A system for shotgun DNA sequencing. Nucleic Acid Res. **9**:309–321.
27. **Michaelis, S., and J. Beckwith.** 1982. Mechanism of incorporation of cell envelope proteins in *Escherichia coli*. Annu. Rev. Microbiol. **36**:435–465.
28. **Miller, J. H.** 1972. Experiments in molecular genetics. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
29. **Noegel, A., U. Rdest, and W. Goebel.** 1981. Determination of the functions of hemolytic plasmid pHly152 of *Escherichia coli*. J. Bacteriol. **145**:233–247.
30. **Noegel, A., U. Rdest, W. Springer, and W. Goebel.** 1979. Plasmid cistrons controlling synthesis and secretion of the exotoxin α-haemolysin of *Escherichia coli*. Mol. Gen. Genet. **175**:343–350.
31. **Perlman, D., and H. O. Halvorson.** 1983. A putative signal peptidase recognition site and sequence in eucaryotic and procaryotic signal peptides. J. Mol. Biol. **167**:391–409.
32. **Pollock, M. R., and M. H. Richmond.** 1962. Low cyst(e)ine content of bacterial extracellular proteins: its possible physiological significance. Nature (London) **194**:446–449.
33. **Pustell, J., and F. C. Kafatos.** 1984. A convenient and adaptable package of computer programs for DNA and protein sequence management, analysis and homology determination. Nucleic Acids Res. **12**:643–655.
34. **Rosenberg, M., and D. Court.** 1979. Regulatory sequences involved in the promotion and termination of RNA transcription. Annu. Rev. Genet. **13**:319–353.
35. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U.S.A. **74**:5463–5467.
36. **Shine, J., and L. Dalgarno.** 1974. The 3′-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. Proc. Natl. Acad. Sci. U.S.A. **71**:1342–1346.
37. **Silhavy, T., S. Benson, and S. Emr.** 1983. Mechanisms of protein localization. Microbiol. Rev. **47**:313–344.
38. **Stark, J. M., and C. W. Shuster.** 1982. Analysis of hemolytic determinants of plasmid pHly185 by Tn*5* mutagenesis. J. Bacteriol. **152**:963–967.
39. **Tinoco, I., P. Borer, B. Dengler, M. Levine, O. Uhlenbeck, D. Crothers, and J. Gralla.** 1973. Improved estimation of secondary structure in ribonucleic acids. Nature (London) New Biol. **246**:40–41.
40. **Viera, J., and J. Messing.** 1982. The pUC plasmids, an M13 mp7-derived system for insertion mutagenesis and sequencing with synthetic universal primers. Gene **19**:259–268.
41. **Von Heijne, G.** 1983. Patterns of amino acids near signal-sequence cleavage sites. Eur. J. Biochem. **133**:17–21.
42. **Wagner, W., M. Vogel, and W. Goebel.** 1983. Transport of hemolysin across the outer membrane of *Escherichia coli* requires two functions. J. Bacteriol. **154**:200–210.
43. **Welch, R. A., E. P. Dellinger, B. Minshew, and S. Falkow.** 1981. Haemolysin contributes to virulence of extraintestinal *Escherichia coli* infections. Nature (London) **294**:665–667.
44. **Welch, R. A., and S. Falkow.** 1984. Characterization of *Escherichia coli* hemolysins conferring quantitative differences in virulence. Infect. Immun. **43**:156–160.
45. **Welch, R. A., R. Hull, and S. Falkow.** 1983. Molecular cloning and physical characterization of a chromosomal hemolysin from *Escherichia coli*. Infect. Immun. **42**:178–186.
46. **Yamamoto, T., T. Tamura, and T. Yokota.** 1983. Primary structure of heat-labile enterotoxin produced by *Escherichia coli* pathogenic for humans. J. Biol. Chem. **259**:5037–5044.