JOURNAL OF BACTERIOLOGY, Dec. 1984, p. 1184–1187
0021-9193/84/121184-04$02.00/0

Vol. 160, No. 3

Nucleotide Sequence of the Gene Encoding Pilin of *Bacteroides nodosus*, the Causal Organism of Ovine Footrot

THOMAS C. ELLEMAN* AND PETER A. HOYNE

CSIRO Division of Protein Chemistry, Parkville, Victoria, 3052, Australia

Received 13 July 1984/Accepted 30 August 1984

The nucleotide sequence encoding pilin, the monomer protein subunit of the pilus from *Bacteroides nodosus*, has been determined. The sequence predicts a short, positively charged, amino-terminal segment which is absent from the amino acid sequence of mature pilin. The coding sequence is preceded upstream by a sequence of five nucleotides complementary to the 3′ end of 16S rRNA of *Escherichia coli*—a potentially good ribosome binding site—and even further upstream by an AT-rich region preceding several potential recognition sites for RNA polymerase. The coding sequence is followed by a region of hyphenated dyad symmetry having the potential to act as a rho-independent terminator of transcription.

Many gram-negative bacteria possess numerous small filamentous surface appendages known as pili (24). Each pilus is composed of identical subunits of the protein pilin. The N-terminal amino acid sequences of pilin from bacterial species as diverse as *Neisseria gonorrhoeae*, *Pseudomonas aeruginosa*, *Moraxella nonliquefaciens*, and *Bacteroides nodosus* show a high degree of sequence similarity (11, 16, 20, 28). Vaccines prepared from purified whole pili have been shown to protect against gonorrhoeal infection by urethral challenge in humans (4), and against footrot in sheep challenged by *B. nodosus* of homologous serogroups (10, 30–32). With a view to the preparation of a pilin vaccine against footrot by a less fastidious organism than the anaerobe *B. nodosus*, we have isolated colonies of *E. coli* RR1 transformed by the fragmented genome of *B. nodosus* (strain 198) cloned in the plasmid vector pBR322 (9). Several transformants produced a protein recognized by anti-pili antibody and of a mobility indistinguishable from that of pilin on sodium dodecyl sulfate-polyacrylamide gel electrophoresis. Mapping studies of these transformants have located the pilin structural gene of *B. nodosus* spanning the single *Pst*I site present in all of the plasmid inserts. This study reports the sequence of that region of the cloned DNA encompassing the coding sequence of the pilin gene.

Sequence determination. Two recombinant clones (5A5 and 7A1E of reference 9) with plasmid inserts in opposite orientations relative to pBR322 were used for sequence determination. The prediction of a short, unusual "signal sequence" in the coding sequence necessitated verification in two unrelated clones. Recombinant plasmid was isolated from bacterial cultures by Triton lysis of lysozyme-EDTA spheroplasts followed by isopycnic centrifugation in cesium chloride-ethidium bromide (18). Large restriction fragments (>600 base pairs) from digestion of plasmids were separated by electrophoresis in 1% agarose gel and recovered by electrophoresis onto NA45 membrane (Schleicher & Schuell, Inc.) (8). DNA sequences were initially determined by "forced" cloning into M13mp8 (21) the *Pst*I-*Eco*RI fragments isolated from clone 5A5, to prepare templates for the dideoxy chain-termination method of sequencing (27). Complementary sequences were verified by the use of synthetic oligonucleotide primers based on the extremities of the determined sequence, with the recombinant plasmid from clones 5A5 and 7A1E as templates. Sequences were also determined for *Sau*3AI-cut fragments prepared from both the isolated *Pst*I-*Eco*RI fragments of clone 5A5 and directly from the smaller recombinant clone 7A1E. These fragments were subcloned in M13mp8, and the sequences were determined by the use of both a universal primer (6) and synthetic primers for sequences distant from the universal priming site (Fig. 1).

Synthetic oligonucleotide primers. Oligonucleotides (5′AACTGACTCTAAAC, 5′CCTACTTAGACAAG, 5′TAAATATCATCTTG) were manually synthesized by a phosphate triester method (7). These primers were purified by ion-exchange high-pressure liquid chromatography on a Radial PAK SAX cartridge (Waters Associates) eluted at a flow rate of 2 ml/min with a gradient of 20 to 240 mM KH_2PO_4 (pH 7.0) containing 30% (vol/vol) ethanol.

Predicted amino acid sequence. The nucleotide sequence spanning the *Pst*I cleavage site of the pBR322-cloned inserts contains a single open reading-frame that encodes the protein sequence of mature pilin preceded by an amino-terminal extension (Fig. 2). Most bacterial proteins destined for secretion or outer membrane insertion are synthesized as precursors that contain additional amino-terminal sequence termed the signal sequence (23). Signal sequences generally consist of a hydrophobic core of 12 to 20 amino acids that is preceded by one or more positively charged residues. The predicted amino acid sequence of pilin lacks a hydrophobic section within the amino-terminal extension, but corresponds to the general overall structure of a presecretory protein since the small additional segment is positively charged and the amino-terminal region of mature pilin is highly hydrophobic. Thus, the short amino-terminal extension together with the amino terminal sequence of mature pilin may constitute a signal sequence. A possible function for this novel structure of presecretory pilin could be to avoid later arrest of transfer across the cytoplasmic membrane by the hydrophobic segment of pilin. Partial translocation in the case of integral membrane proteins destined for membrane insertion rather than secretion is attained by arresting transfer at a highly hydrophobic segment after the signal sequence (35). Inclusion of the highly hydrophobic N-terminal sequence of mature pilin within the signal sequence would serve to prevent this sequence from acting as a "stop-transfer" sequence (2).

Cleavage of a signal sequence is not a prerequisite for export (19), although signal sequences are generally proteolytically removed from presecretory proteins. Small residues

* Corresponding author.

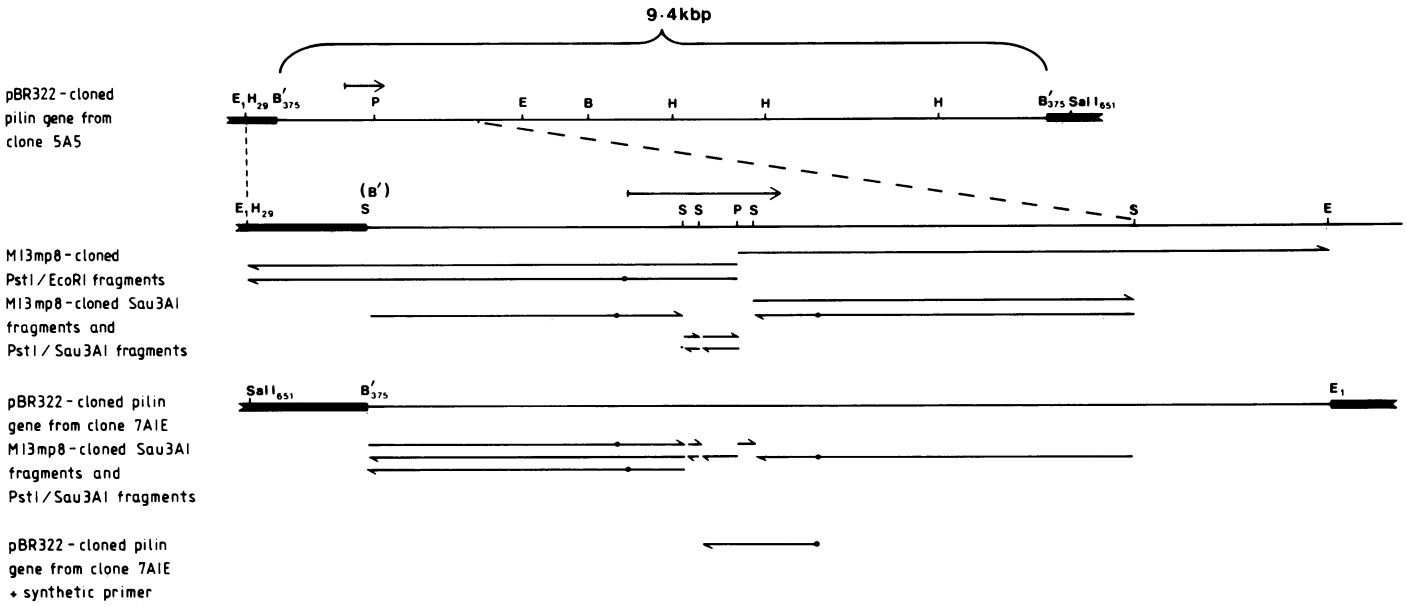

FIG. 1. Summary of sequencing strategy. The pBR322-cloned inserts from clones 5A5 and 7A1E, and their derived restriction fragments cloned in M13mp8 for use in sequence determination, are shown. The pBR322 vector is indicated by thick lines. Restriction enzyme sites for *Eco*RI (E), *Hin*dIII (H), *Bam*HI (B), *Pst*I (P), and *Sal*I are indicated, together with the *Bam*HI site (B′) of pBR322 into which the DNA of *B. nodosus* was inserted. *Sau*3AI sites (S) are only indicated within that section of the insert DNA shown expanded between the dotted lines. Locations of synthetic primer sites are indicated by dots. The predicted coding region and direction of transcription of pilin is indicated by arrows.

are found preceding the cleavage point, and in this respect the presecretory pilin sequence resembles other signal sequences since the amino acid preceding the mature sequence of pilin is glycine. The ultimate function of the N-terminal hydrophobic segment in mature pilin may be to form the inter-subunit contacts between pilin monomers within the pilus. That high precision is required, whatever the function, is suggested by the outstanding sequence conservation (100%) seen between the hydrophobic N-terminal segments of pilin from diverse genera of bacteria. Sequence homology is not apparent elsewhere in the molecules (20).

Several differences occur between the sequence of pilin predicted from the gene and the reported amino acid sequence (20). The disparities at amino acid residues 73, 95, and 128 have since been recognized as errors in the determination of the amino acid sequence (N. M. McKern, personal communication). The only other disparity involves the assignment of an amide group to aspartic acid in the determined amino acid sequence, the gene sequence predicting aspartic acid at position 87 of mature pilin from clones 5A5 and 7A1E.

Translation and transcription signals. The only potential translation initiation codons (AUG, GUG, AUA, UUG) found within the open reading frame preceding the coding sequence of mature pilin occur in the $_{153}$ATGATG sequence. A potential Shine-Dalgarno sequence (29), $_{145}$AGGAG, occurring three nucleotides upstream from this sequence, suggests that translation initiation occurs at the second ATG triplet, since the distance between the proposed Shine-Dalgarno sequence and the first ATG triplet is much shorter than the minimal five bases generally found for initiation codons (33). This first triplet also lacks other markers characteristic of initiation codons (12, 33), whereas the second ATG triplet has several. These markers are (i) AAAA immediately after the ATG triplet, and (ii) TA eight residues after the ATG triplet, both common sequences at these positions relative to initiation codons. Termination codons overlap or closely precede the translation initiation site in all three reading frames. Termination codons in this region are characteristic of many bacterial initiation sites (1), although this may only reflect the high incidence of A in this region (12).

After the protein coding sequence of the pilin gene are two adjacent termination codons. The former, TAA, is thought to be the major termination codon used in *Escherichia coli* (3). These are followed closely downstream by a region of hyphenated dyad symmetry ending in a run of T residues. A very stable RNA stem and loop structure ($\Delta G = -18.4$ kilocalories [77.0 kJ] [34]) can be formed from this sequence. This feature is characteristic of the 3′ termini of many bacterial mRNAs, indicating a potential rho-independent transcription termination signal for RNA polymerase (26).

Without transcript mapping, promoter sequences are difficult to identify. Homology with the canonical recognition site (−35 site) and binding site (−10 site) of *E. coli* RNA polymerase (15, 26) is often so low in many identified promoters that their identification would be impossible without prior knowledge of transcription initiation sites. A feature of many strong promoters (5, 22, 25), however, is an extensive AT-rich region preceding the recognition site and believed to lower the melting temperature of the DNA, thus facilitating RNA polymerase-mediated strand unwinding. Such an outstanding AT-rich region (nucleotides 21 to 63 in Fig. 2) occurs upstream of the pilin coding sequence. Within and after this putative entry point for RNA polymerase are several sites showing similarity with the canonical TTGACA recognition site and the TATAAT binding site of RNA polymerase. A short region of hyphenated dyad symmetry, a feature sometimes found associated with the RNA polymerase binding site, is situated a few nucleotides upstream of the potential ribosome binding site.

Codon usage. It is interesting to consider the codon usage

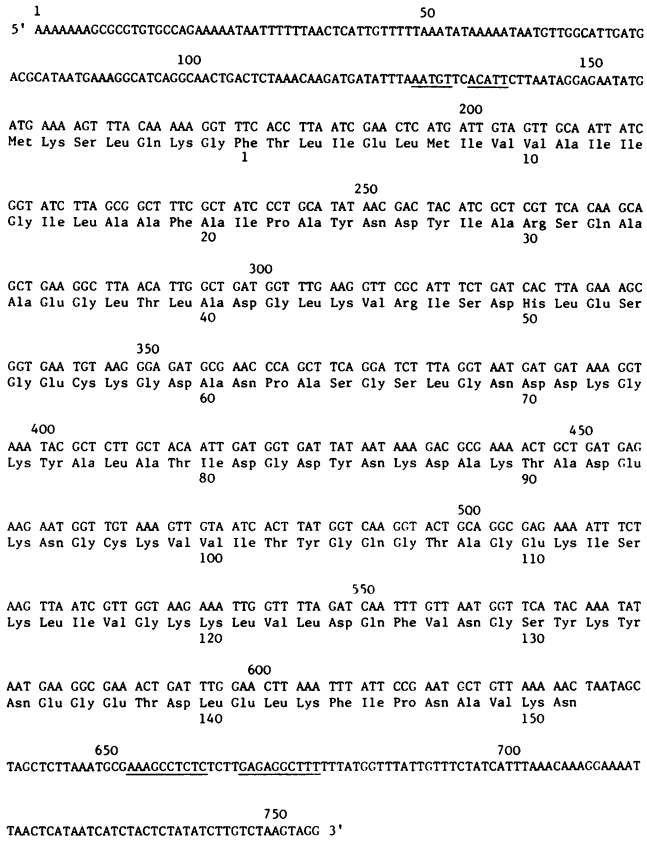

FIG. 2. Nucleotide sequence encoding pilin of *B. nodosus* (strain 198). The sequence was derived independently from clones 5A5 and 7A1E. Only that DNA strand corresponding to the mRNA sequence is shown. The predicted amino acid sequence of the open reading frame containing the coding sequence of pilin is numbered from the first amino acid of mature pilin. Two regions of hyphenated dyad symmetry are underlined.

of the pilin gene from *B. nodosus* (Table 1) in relation to the codon usage of *E. coli* since this has an important bearing on the expression of pilin by this foreign host. In *E. coli*, the genes of proteins produced in large quantities show a highly preferential use of certain codons recognized by the most abundant isoaccepting species of tRNA. Codons recognized by the minor tRNA species are rarely used (14). These latter codons (AUA, CUA, CG^{A}_{G}, AG^{A}_{G}, GG^{A}_{G}), which might result in inefficient translation if used excessively in a foreign gene, are largely absent in the pilin coding sequence (Table 1). Other codons used only sparingly in the highly expressed genes of *E. coli* (UC^{A}_{G}, ACG, CCC [13]) are also infrequent in the pilin coding sequence.

The pilin coding sequence shows a highly preferential use of T in the third position of the codon. This use is most predominant in the quartet codons. A similarly preferential use of T as the third base in the quartet codons has been recognized in the highly expressed genes of *E. coli*. It has been suggested that the preferential use of T rather than C as the third base of the codon facilitates the efficient translation of highly expressed *E. coli* genes by averting the formation of the most tightly bound codon-anticodon complexes (13). This feature should also facilitate expression of pilin in *E. coli*.

The major disparity between the overall codon usage in genes of *E. coli* and that of the pilin gene is the lack of use of the CTG codon for leucine in the latter. In *E. coli*, 50% of all occurrences of leucine are coded by CTG. However, tRNA molecules interacting with CU^{C}_{U} or UU^{A}_{G} are abundant in *E. coli* (17) and should allow efficient translation of all the leucine codons in the pilin gene.

TABLE 1. Frequency of codon usage in the predicted coding region of the pilin gene[a]

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

[a] Possible initiation codons are excluded. Codons corresponding to minor (or weakly interacting) tRNA species in *E. coli* are underlined. First position totals: U, 31; C, 13; A, 49; G, 64. Second position totals: U, 42; C, 33; A, 59; G, 23. Third position totals: U, 71; C, 25; A, 45; G, 16.

The nucleotide sequence determined for the coding region of pilin, together with the neighboring DNA sequence, suggests that it may be feasible to express in *E. coli* the large quantities of pilin necessary for the production of a single protein vaccine to replace the conventional whole-cell *B. nodosus* vaccine. The codon usage, the ribosome binding site, and the transcription termination site all resemble their counterparts in *E. coli* genes. Only the putative promoter region and the unusual signal sequence are without a well-defined equivalent in *E. coli* and might require genetic manipulation for efficient expression in this host. Amino acid sequence studies (N. McKern, personal communication) on the protein as expressed in *E. coli* RR1 from a strong phage λ promoter have indicated that the short signal sequence predicted by the gene sequence is not removed in this organism.

The whole-cell *B. nodosus* vaccine currently in use is expensive to manufacture because of the fastidious nature of the organism and its sparse growth and variable pilus expression when grown in bulk liquid culture. In addition, the large amount of extraneous antigen in this vaccine causes vaccinal lesions (31). A single protein vaccine produced by recombinant DNA techniques could circumvent all of these problems of the conventional vaccine.

We are grateful to D. J. Stewart, B. L. Clark, and D. L. Emery for introducing this problem to us, and for their collaboration in the cloning and identification of *E. coli* recombinants containing the pilin gene of *B. nodosus*.

LITERATURE CITED

1. **Atkins, J. F.** 1979. Is UAA or UGA part of the recognition signal for ribosomal initiation? Nucleic Acids. Res. **7**:1035–1041.
2. **Blobel, G.** 1980. Intracellular protein topogenesis. Proc. Natl. Acad. Sci. U.S.A. **77**:1496–1500.
3. **Brenner, S., A. O. Stretton, and S. Kaplan.** 1965. Genetic code: the 'nonsense' triplets for chain termination and their suppression. Nature (London) **206**:994–998.

4. **Brinton, C. C., Jr., S. W. Wood, A. Brown, A. M. Labik, J. R. Bryan, S. W. Lee, S. E. Polen, E. C. Tramont, J. Sandoff, and W. Zollinger.** 1982. The development of a neisserial pilus vaccine for gonorrhea and meningococcal meningitis. Semin. Infect. Dis. **4**:140–159.
5. **De Boer, H. A., S. F. Gilbert, and M. Nomura.** 1979. DNA sequences of promoter regions for rRNA operons rrnE and rrnA in *E. coli*. Cell **17**:201–209.
6. **Duckworth, M. L., M. J. Gait, P. Goelet, G. F. Hong, M. Singh, and R. C. Titmas.** 1981. Rapid synthesis of oligodeoxyribonucleotides. VI. Efficient, mechanized synthesis of heptadecadeoxyribonucleotides by an improved solid phase phosphotriester route. Nucleic Acids Res. **9**:1691–1706.
7. **Efimov, V. A., S. V. Reverdatto, and O. G. Chakhmakhcheva.** 1982. New effective method for the synthesis of oligonucleotides via phosphotriester intermediates. Nucleic Acids Res. **10**:6675–6694.
8. **Elleman, T. C., P. A. Hoyne, M. L. Dyall-Smith, I. H. Holmes, and A. A. Azad.** 1983. Nucleotide sequence of the gene encoding the serotype-specific glycoprotein of UK bovine rotavirus. Nucleic Acids Res. **11**:4689–4701.
9. **Elleman, T. C., P. A. Hoyne, D. L. Emery, D. J. Stewart, and B. L. Clark.** 1984. Isolation of the gene encoding pilin of *Bacteroides nodosus* (strain 198), the causal organism of ovine footrot. FEBS Lett. **173**:103–107.
10. **Every, D., and T. M. Skerman.** 1982. Protection of sheep against experimental footrot by vaccination with pili purified from *Bacteroides nodosus*. N.Z. Vet. J. **30**:156–158.
11. **Frøholm, L. O., and K. Sletten.** 1977. Purification and N-terminal sequence of a fimbrial protein from *Moraxella nonliquefaciens*. FEBS Lett. **73**:29–32.
12. **Gold, L., D. Pribnow, T. Schneider, S. Shinedling, B. S. Singer, and G. D. Stormo.** 1981. Translational initiation in prokaryotes. Annu. Rev. Microbiol. **35**:365–403.
13. **Grantham, R., C. Gautier, M. Gouy, M. Jaconzone, and R. Mercier.** 1981. Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res. **9**:r43–r74.
14. **Grosjean, H., and W. Fiers.** 1982. Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. Gene **18**:199–209.
15. **Hawley, D. K., and W. R. McClure.** 1983. Compilation and analysis of *Escherichia coli* promoter DNA sequences. Nucleic Acids Res. **11**:2237–2255.
16. **Hermodson, M. A., K. C. S. Chen, and T. M. Buchanan.** 1978. *Neisseria* pili proteins: amino-terminal amino acid sequences and identification of an unusual amino acid. Biochemistry **17**:442–445.
17. **Ikemura, T.** 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. J. Mol. Biol. **146**:1–21.
18. **Kupersztoch, Y. M., and D. R. Helinski.** 1973. A catenated DNA molecule as an intermediate in the replication of the resistance transfer factor R6K in *Escherichia coli*. Biochem. Biophys. Res. Commun. **54**:1451–1459.
19. **Lin, J. J. C., H. Kanazawa, J. Ozols, and H. C. Wu.** 1978. An *Escherichia coli* mutant with an amino acid alteration within the signal sequence of outer membrane prolipoprotein. Proc. Natl. Acad. Sci. U.S.A. **75**:4891–4895.
20. **McKern, N. M., I. J. O'Donnell, A. S. Inglis, D. J. Stewart, and B. L. Clark.** 1983. Amino acid sequence of pilin from *Bacteroides nodosus* (strain 198), the causative organism of ovine footrot. FEBS Lett. **164**:149–153.
21. **Messing, J.** 1983. New M13 vectors for cloning. Methods Enzymol. **101**:20–78.
22. **Nakamura, K., and I. Massayori.** 1979. DNA sequence of the gene for the outer membrane lipoprotein of *E. coli*: an extremely AT-rich promoter. Cell **18**:1109–1117.
23. **Pages, J. M.** 1983. Biosynthése et exportation des protéines de l'enveloppe d'*Escherichia coli*. Biochemie **65**:531–541.
24. **Pearce, W. A., and T. M. Buchanan.** 1980. Structure and cell membrane-binding properties of bacterial fimbriae, p. 291–344. *In* E. H. Beachey (ed.), Receptors and recognition. Bacterial adherence, series B, vol. 6. Chapman and Hall, London.
25. **Post, L. E., A. E. Arfsten, G. R. Davies, and M. Nomura.** 1980. DNA sequence of the promoter region for the α ribosomal protein operon in *Escherichia coli*. J. Biol. Chem. **255**:4653–4659.
26. **Rosenburgh, M., and D. Court.** 1979. Regulatory sequences involved in the promotion and termination of RNA transcription. Annu. Rev. Genet. **13**:319–353.
27. **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.
28. **Sastry, P. A., J. R. Pearlstone, L. B. Smillie, and W. Paranchych.** 1983. Amino acid sequence of pilin isolated from *Pseudomonas aeruginosa* PAK. FEBS Lett. **151**:253–256.
29. **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.
30. **Stewart, D. J.** 1978. The role of various antigenic fractions of *Bacteroides nodosus* in eliciting protection against foot-rot in vaccinated sheep. Res. Vet. Sci. **24**:14–19.
31. **Stewart, D. J., B. L. Clark, D. L. Emery, J. E. Peterson, and K. J. Fahey.** 1983. A *Bacteroides nodosus* immunogen, distinct from the pilus, which induces cross-protective immunity in sheep vaccinated against footrot. Aust. Vet. J. **60**:83–84.
32. **Stewart, D. J., B. L. Clark, J. E. Peterson, D. A. Griffiths, and E. F. Smith.** 1982. Importance of pilus-associated antigen in *Bacteroides nodosus* vaccines. Res. Vet. Sci. **32**:140–147.
33. **Stormo, G. D., T. D. Schneider, and L. M. Gold.** 1982. Characterization of translational initiation sites in *E. coli*. Nucleic Acids Res. **10**:2971–2996.
34. **Tinoco, I., P. N. Borer, B. Dengler, M. D. Levine, O. C. Uhlenbeck, D. M. Crothers, and J. Gralla.** 1973. Improved estimation of secondary structure in ribonucleic acids. Nature (London) New Biol. **246**:40–41.
35. **Yost, C. S., J. Hedgpeth, and V. R. Lingappa.** 1983. A stop transfer sequence confers predictable transmembrane orientation to a previously secreted protein in cell-free systems. Cell **34**:759–766.