JOURNAL OF BACTERIOLOGY, Dec. 1986, p. 1479–1483
0021-9193/86/121479-05$02.00/0

Vol. 168, No. 3

Nucleotide Sequence of the *hag* Gene Encoding Flagellin of *Escherichia coli*

GORO KUWAJIMA,* JUN-ICHIRO ASAKA, TAMIO FUJIWARA, TAKASHI FUJIWARA, KAZUMI NODE, AND EIJI KONDO

Shionogi Research Laboratories, Shionogi & Co., Ltd., Fukushima-ku, Osaka 553, Japan

Received 24 April 1986/Accepted 2 September 1986

We determined the DNA sequence of the *hag* gene of *Escherichia coli* K-12 and deduced the primary structure of the flagellin consisting of 497 amino acid residues. Comparison of the amino acid sequence with those of other bacterial flagellins revealed a high homology in the NH_2- and COOH-terminal regions.

Flagellin is the subunit protein which polymerizes to form filaments of bacterial flagella. The flagellum apparatus of *Salmonella typhimurium* and *Escherichia coli* have been the subject of extensive genetic and physicochemical studies (5, 6). However, at the start of this work, the complete amino acid sequences of the flagellins in these bacteria remained to be clarified, and only the sequence of about 60 nucleotides had been reported for the DNA sequences of the structural genes encoding the flagellins of *S. typhimurium* (*H*1 and *H*2) and *E. coli* (*hag*) (20). Here we report the complete DNA sequence of the *E. coli* K-12 *hag* gene and the primary structure of the flagellin deduced from it. Meanwhile, the DNA sequences of the genes encoding the flagellins of four *Salmonella* spp. were also reported (8, 21). Therefore, we could compare the amino acid sequences of the flagellins from *E. coli* and *S. typhimurium* to reveal a high homology in the NH_2- and COOH-terminal regions but not in the central region.

The *hag* gene of *E. coli* K-12 had been previously cloned on phage λ p*fla*H_2 (10). The 7.5-kiolobase-pair (kbp) *Eco*RI-

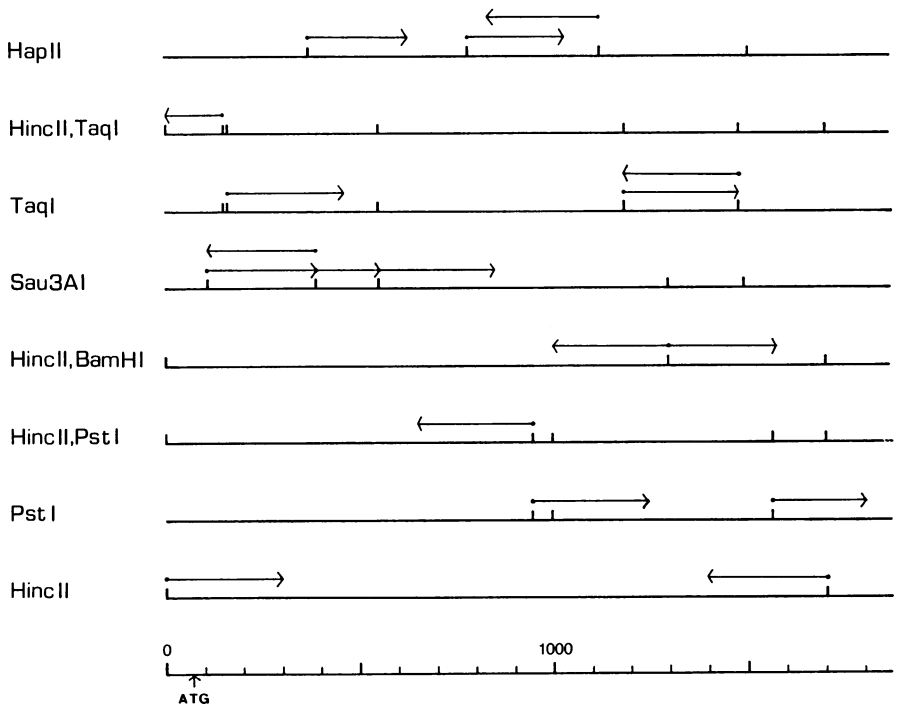

FIG. 1. Restriction nuclease digestion map of the *Hin*cII fragment containing the *hag* gene and sequence strategy. Arrows indicate the extent of the determined DNA sequence and are aligned in the 5′ → 3′ direction. The location of the initiation codon ATG is indicated. Scale, nucleotide number from 1 of the *Hin*cII site.

* Corresponding author.

```
                                                                GACGGCGAT

TGAGCCGACGGGTGGAAACCCAATACGTAATCAACGACTTGCAATATAGGATAACGAATC

         10         20         30         40         50         60
ATGGCACAAGTCATTAATACCAACAGCCTCTCGCTGATCACTCAAAATAATATCAACAAG
MetAlaGlnValIleAsnThrAsnSerLeuSerLeuIleThrGlnAsnAsnIleAsnLys

         70         80         90        100        110        120
AACCAGTCTGCGCTGTCGAGTTCTATCGAGCGTCTGTCTTCTGGCTTGCGTATTAACAGC
AsnGlnSerAlaLeuSerSerSerIleGluArgLeuSerSerGlyLeuArgIleAsnSer

        130        140        150        160        170        180
GCGAAGGATGACGCAGCGGGTCAGGCGATTGCTAACCGTTTCACCTCTAACATTAAAGGC
AlaLysAspAspAlaAlaGlyGlnAlaIleAlaAsnArgPheThrSerAsnIleLysGly

        190        200        210        220        230        240
CTGACTCAGGCGGCCCGTAACGCCAACGACGGTATCTCCGTTGCGCAGACCACCGAAGGC
LeuThrGlnAlaAlaArgAsnAlaAsnAspGlyIleSerValAlaGlnThrThrGluGly

        250        260        270        280        290        300
GCGCTGTCCGAAATCAACAACAACTTACAGCGTGTGCGTGAACTGACGGTACAGGCCACT
AlaLeuSerGluIleAsnAsnAsnLeuGlnArgValArgGluLeuThrValGlnAlaThr

        310        320        330        340        350        360
ACCGGTACTAACTCTGAGTCTGATCTGTCTTCTATCCAGGACGAAATTAAATCCCGTCTG
ThrGlyThrAsnSerGluSerAspLeuSerSerIleGlnAspGluIleLysSerArgLeu

        370        380        390        400        410        420
GATGAAATTGACCGCGTATCTGGTCAGACCCAGTTCAACGGCGTGAACGTGCTGGCAAAA
AspGluIleAspArgValSerGlyGlnThrGlnPheAsnGlyValAsnValLeuAlaLys

        430        440        450        460        470        480
AATGGCTCCATGAAAATCCAGGTTGGCGCAAATGATAACCAGACTATCACTATCGATCTG
AsnGlySerMetLysIleGlnValGlyAlaAsnAspAsnGlnThrIleThrIleAspLeu

        490        500        510        520        530        540
AAGCAGATTGATGCTAAAACTCTTGGCCTTGATGGTTTTAGCGTTAAAAATAACGATACA
LysGlnIleAspAlaLysThrLeuGlyLeuAspGlyPheSerValLysAsnAsnAspThr

        550        560        570        580        590        600
GTTACCACTAGTGCTCCAGTAACTGCTTTTGGTGCTACCACCACAAACAATATTAAACTT
ValThrThrSerAlaProValThrAlaPheGlyAlaThrThrThrAsnAsnIleLysLeu

        610        620        630        640        650        660
ACTGGAATTACCCTTTCTACGGAAGCAGCCACTGATACTGGCGGAACTAACCCAGCTTCA
ThrGlyIleThrLeuSerThrGluAlaAlaThrAspThrGlyGlyThrAsnProAlaSer

        670        680        690        700        710        720
ATTGAGGGTGTTTATACTGATAATGGTAATGATTACTATGCGAAAATCACCGGTGGTGAT
IleGluGlyValTyrThrAspAsnGlyAsnAspTyrTyrAlaLysIleThrGlyGlyAsp

        730        740        750        760        770        780
AACGATGGGAAGTATTACGCAGTAACAGTTGCTAATGATGGTACAGTGACAATGGCGACT
AsnAspGlyLysTyrTyrAlaValThrValAlaAsnAspGlyThrValThrMetAlaThr

        790        800        810        820        830        840
GGAGCAACGGCAAATGCAACTGTAACTGATGCAAATACTACTAAAGCTACAACTATCACT
GlyAlaThrAlaAsnAlaThrValThrAspAlaAsnThrThrLysAlaThrThrIleThr
```

FIG. 2. DNA nucleotide sequence of *hag* and amino acid sequence of flagellin. The first letter A of the translational initiation codon is nucleotide 1. The underlined AGGA is considered to be the ribosome-binding site. The DNA sequence of λ *c*I857 used to clone *hag* is indicated with a broken line.

```
      850        860        870        880        890        900
TCAGGCGGTACACCTGTTCAGATTGATAATACTGCAGGTTCCGCAACTGCCAACCTTGGT
SerGlyGlyThrProValGlnIleAspAsnThrAlaGlySerAlaThrAlaAsnLeuGly

      910        920        930        940        950        960
GCTGTTAGCTTAGTAAAACTGCAGGATTCCAAGGGTAATGATACCGATACATATGCGCTT
AlaValSerLeuValLysLeuGlnAspSerLysGlyAsnAspThrAspThrTyrAlaLeu

      970        980        990       1000       1010       1020
AAAGATACAAATGGCAATCTTTACGCTGCGGATGTGAATGAAACTACTGGTGCTGTTTCT
LysAspThrAsnGlyAsnLeuTyrAlaAlaAspValAsnGluThrThrGlyAlaValSer

     1030       1040       1050       1060       1070       1080
GTTAAAACTATTACCTATACTGACTCTTCCGGTGCCGCCAGTTCTCCAACCGCGGTCAAA
ValLysThrIleThrTyrThrAspSerSerGlyAlaAlaSerSerProThrAlaValLys

     1090       1100       1110       1120       1130       1140
CTGGGCGGAGATGATGGCAAAACAGAAGTGGTCGATATTGATGGTAAAACATACGATTCT
LeuGlyGlyAspAspGlyLysThrGluValValAspIleAspGlyLysThrTyrAspSer

     1150       1160       1170       1180       1190       1200
GCCGATTTAAATGGCGGTAATCTGCAAACAGGTTTGACTGCTGGTGGTGAGGCTCTGACT
AlaAspLeuAsnGlyGlyAsnLeuGlnThrGlyLeuThrAlaGlyGlyGluAlaLeuThr

     1210       1220       1230       1240       1250       1260
GCTGTTGCAAATGGTAAAACCACGGATCCGCTGAAAGCGCTGGACGATGCTATCGCATCT
AlaValAlaAsnGlyLysThrThrAspProLeuLysAlaLeuAspAspAlaIleAlaSer

     1270       1280       1290       1300       1310       1320
GTAGACAAATTCCGTTCTTCCCTCGGTGCGGTGCAAAACCGTCTGGATTCCGCGGTTACC
ValAspLysPheArgSerSerLeuGlyAlaValGlnAsnArgLeuAspSerAlaValThr

     1330       1340       1350       1360       1370       1380
AACCTGAACAACACCACTACCAACCTGTCTGAAGCGCAGTCCCGTATTCAGGACGCCGAC
AsnLeuAsnAsnThrThrThrAsnLeuSerGluAlaGlnSerArgIleGlnAspAlaAsp

     1390       1400       1410       1420       1430       1440
TATGCGACCGAAGTGTCCAATATGTCGAAAGCGCAGATCATCCAGCAGGCCGGTAACTCC
TyrAlaThrGluValSerAsnMetSerLysAlaGlnIleIleGlnGlnAlaGlyAsnSer

     1450       1460       1470       1480       1490
GTGTTGGCAAAAGCTAACCAGGTACCGCAGCAGGTTCTGTCTCTGCTGCAGGGTTAATCG
ValLeuAlaLysAlaAsnGlnValProGlnGlnValLeuSerLeuLeuGlnGly***

TTGTAACCTGATTAACTGAGACTGACGGCAACGCCAAATTGCCTGATGCGCTGCGCTTAT

CAGGCCTACAAGTTGAATTGCAATTTATTGAATTTGCACCCCAGGCCAGTGCTTTAGCGT

T
```

*Sal*I, fragment of λ p*fla*H_2 DNA was inserted into plasmid pBR322 (1). Then, the 3.5-kbp *Bam*HI fragment from λ *H2* was deleted, and plasmid pBR322/hag93 was obtained. As pBR322/hag93 could confer motility to *E. coli* W3623H *fla-am*76, which carries an amber mutation in the *hag* gene (12, 13), it was confirmed that pBR322/hag93 carried the *hag* gene. Comparison of the restriction enzyme cleavage map of pBR322/hag93 with those of the upstream region of the *hag* gene (20) and λ phage DNA suggested that the protein-coding region of the *hag* gene was in the 1.7-kbp *Hin*cII fragment of pBR322/hag93. Then, we determined the entire DNA sequence of the 1.7-kbp fragment. The sequence strategy is shown in Fig. 1. DNA fragments from pBR322/hag93 digested with various restriction nucleases were sequenced by the dideoxynucleotide method (17) by using bacteriophages M13mp8, M13mp9, and M13mp18 (15, 22). The DNA sequence of the 1.7-kbp *Hin*cII fragment is presented in Fig. 2. The first 129 nucleotides coincide with

TABLE 1. Amino acid composition of *E. coli* K-12 flagellin

Amino acid	No. of residues/molecule	
	Amino acid analysis[a]	DNA sequence
Ala	58.1	59
Val	32.8[b]	33
Leu	37.5	37
Ile	27.0[b]	28
Gly	45.3	44
Pro	6.0	6
Cys	0.0	0
Met	2.7	3
His	0.0	0
Phe	5.0	5
Tyr	9.9	10
Trp	0.0	0
Asn (+Asp)	88.7	48 (87)
Gln (+Glu)	41.5	27 (41)
Ser	42.8[c]	43
Thr	63.3[c]	65
Lys	25.4	25
Arg	10.2	11

[a] Calculated by assuming the number of Phe residues to be 5.0.
[b] Values are from a 72-h hydrolysate.
[c] Corrected for destruction during hydrolysis.

the results of Szekely and Simon (20). The last 23 bases agree with the DNA sequence of λ phage DNA (31787 to 31809 [16]). The longest translational open reading frame is found from nucleotides 1 to 1497. This frame is preceded by a typical ribosome-binding sequence (AGGA). Other possible open reading frames are too short to encode flagellin.

To confirm that the open reading frame from nucleotides 1 to 1497 was the protein-coding region of the *hag* gene, we analyzed the amino acid composition and the amino acid sequences of both ends of purified flagellin. Flagellin of *E. coli* W3110 was prepared as previously described (11) and further purified by DEAE-cellulose column chromatography. The amino acid composition of the flagellin was analyzed with a Hitachi model 835 amino acid analyzer after hydrolysis of purified flagellin with methanesulfonic acid (19) (Table 1). The NH_2 terminus of the flagellin was also determined to be NH_2-Ala-Glx (Gln or Glu) by the Edman method (7). The COOH terminus of the flagellin was digested with carboxypeptidase P, and the released amino acids were analyzed (23). After 6 h of digestion 1.0 mol each of Gly, Ser, and Val, 1.4 mol of Gln (or Glu), and 2.8 mol of Leu per 1.0 mol of flagellin were detected. Gly was released faster than Ser and Gln (or Glu), and Ser was released faster than Val. Thus, the amino acid composition and the NH_2- and COOH-terminal sequences of mature flagellin agreed well with those deduced from the DNA sequence. Based on these findings,

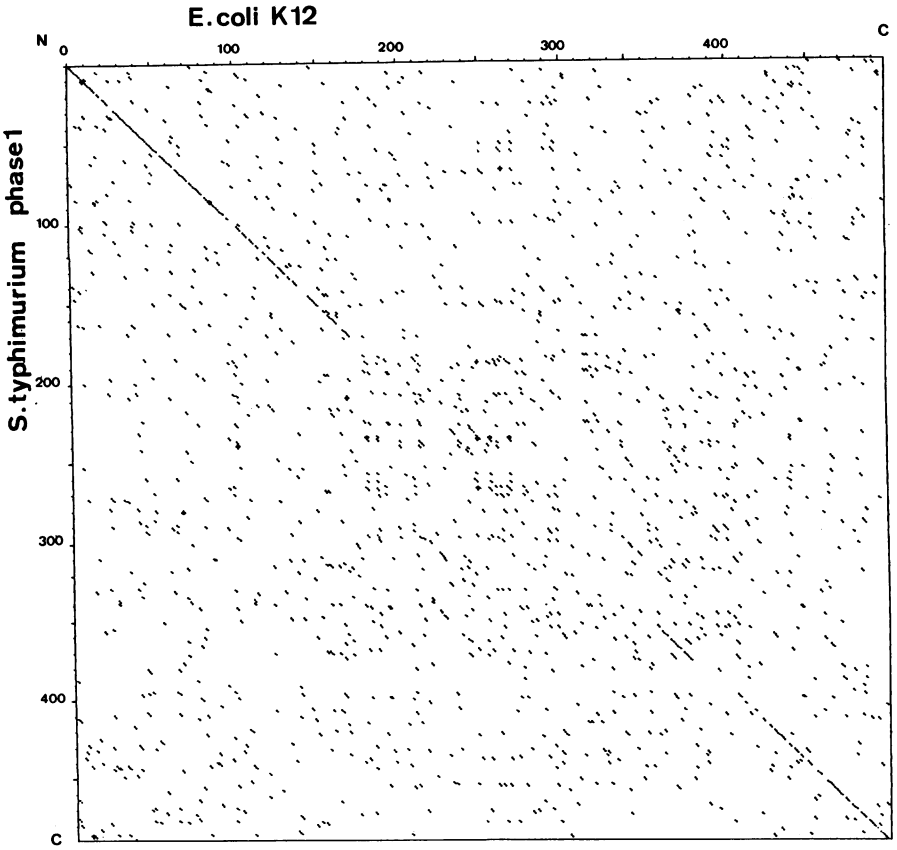

FIG. 3. Dot matrix comparison of the amino acid sequences of *E. coli* K-12 flagellin (horizontal axis) and *S. typhimurium* phase 1 flagellin (vertical axis). The numbers in both axes correspond to the residue numbers from the NH_2 terminus. The points at which at least two consecutive amino acid residues of the ordinate coincide with those of the abscissa are indicated by dots. N and C, Nh_2 and COOH terminus of the flagellin, respectively.

we conclude that the longest open reading frame of nucleotides 1 to 1497 is the protein-coding region of the *hag* gene.

According to our results, *E. coli* K-12 flagellin is composed of 497 amino acid residues, and its molecular weight is 51,172. The amino acid composition (Table 1) shows that the flagellin contains abundant Ala, Val, Leu, Ile, Gly, Ser, Thr, Asn, Asp, Gln, and Lys. These 11 residues compose more than 90% of the total amino acid residues. No Cys, His, and Trp are present. In addition, there are also 53 acidic amino acid residues, in contrast to 36 basic ones (Table 1). Therefore, the flagellin seems to be an acidic protein. Similar characteristics of the amino acid composition, except for the content of His, are common to bacterial flagellins of not only *E. coli* but also *Salmonella* spp., *Bacillus* spp., and other bacteria (3, 4, 9, 14, 18). As far as clarified, the absence of His in flagellin is unique to *E. coli*; other bacterial flagellins have a few His residues.

Until now, the primary structures of flagellins have been clarified for *Bacillus subtilis*, *Caulobacter crescentus*, and four *Salmonella* spp. (2, 3, 8, 21). Among the amino acid sequences of these flagellins, a high homology is seen in the NH_2-terminal and COOH-terminal regions but not in the central region. The flagellin of *E. coli* K-12 also has two terminal regions homologous with those of the flagellins of the above mentioned bacteria, especially *S. typhimurium*. A comparison of *E. coli* K-12 flagellin and *S. typhimurium* phase 1 flagellin allowed us to divide the flagellins into three regions, the NH_2-terminal region of 170 residues, the COOH-terminal region of 140 residues, and the central region of 190 residues (Fig. 3). The extent of homology is about 80% in the NH_2-terminal region, about 60% in the COOH-terminal region, and about 20% in the central region. Joys and Wei supposed that these homologous and heterologous regions in a flagellin molecule were related to its functions, such as its migration to the top of the flagellar hook, its polymerization, and H antigenicity (8, 21). Hereafter in the study of the flagellins it will be necessary to correlate these functions to the amino acid sequences in detail. Our results must be useful to such studies.

We thank Haruo Ozeki for his helpful discussion and advice in preparing this article and Hisato Kondoh for providing phage and bacterial strains.

LITERATURE CITED

1. **Bolivar, F., R. L. Rodriguez, P. J. Greene, M. C. Betlach, H. L. Heyneker, and H. W. Boyer.** 1977. Construction and characterization of new cloning vehicles. II. A multipurpose cloning system. Gene **2**:95–113.
2. **DeLange, R. J., J. Y. Chang, J. H. Shaper, and A. N. Glazer.** 1976. Amino acid sequence of flagellin of *Bacillus subtilis* 168. III. Tryptic peptides, N-bromosuccinimide peptides, and the complete amino acid sequence. J. Biol. Chem. **251**:705–711.
3. **Gill, P. R., and N. Agabian.** 1983. The nucleotide sequence of the M_r = 28,500 flagellin gene of *Caulobacter crescentus*. J. Biol. Chem. **258**:7395–7401.
4. **Gufffanti, A. A., and H. C. Eisenstein.** 1983. Purification of flagella from the alkalophile *Bacillus firmus* RAB. J. Gen. Microbiol. **129**:3239–3242.
5. **Iino, T.** 1969. Genetics and chemistry of bacterial flagella. Bacteriol. Rev. **33**;454–475.
6. **Iino, T.** 1977. Genetics of structure and function of bacterial flagella. Annu. Rev. Genet. **11**:161–182.
7. **Iwanaga, S., P. Wallen, N. J. Grondahl, A. Henschen, and B. Blomback.** 1969. On the primary structure of human fibrinogen: isolation and characterization of N-terminal fragments from plasmic digests. Eur. J. Biochem. **8**:189–199.
8. **Joys, T. M.** 1985. The covalent structure of the phase-1 flagellar filament protein of *Salmonella typhimurium* and its comparison with other flagellins. J. Biol. Chem. **260**:15758–15761.
9. **Joys, T. M., and V. Rankis.** 1972. The primary structure of the phase-1 flagellar protein of *Salmonella typhimurium*. I. The tryptic peptides. J. Biol. Chem. **247**:5180–5193.
10. **Kondoh, H.** 1977. Isolation and characterization of nondefective transducing lambda bacteriophages carrying *fla* genes of *Escherichia coli* K-12. J. Bacteriol. **130**:736–745.
11. **Kondoh, H., and H. Hotani.** 1974. Flagellin from *Escherichia coli* K-12: polymerization and molecular weight in comparison with *Salmonella* flagellins. Biochim. Biophys. Acta. **336**: 117–139.
12. **Kondoh, H., and H. Ozeki.** 1976. Deletion and amber mutation of *fla* loci in *Escherichia coli* K-12. Genetics **84**:403–421.
13. **Kondoh, H., and H. Ozeki.** 1981. Two classes of region III flagellar genes in *Escherichia coli*. J. Bacteriol. **146**:823–825.
14. **McDonough, M. W.** 1965. Amino acid composition of antigenically distinct *Salmonella* flagellar proteins. J. Mol. Biol. **12**: 342–355.
15. **Messing, J.** 1983. New M13 vectors for cloning. Methods Enzymol. **101**:20–78.
16. **Sanger, F., A. R. Coulson, G. F. Hong, D. F. Hill, and G. B. Petersen.** 1982. Nucleotide sequence of bacteriophage λDNA. J. Mol. Biol. **162**:729–773.
17. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74**:5463–5467.
18. **Simon, M. I., S. U. Emerson, J. H. Shaper, P. D. Bernard, and A. N. Glazer.** 1977. Classification of *Bacillus subtilis* flagellins. J. Bacteriol. **130**:200–204.
19. **Simpson, R. J., M. R. Neuberger, and T. Y. Liu.** 1976. Complete amino acid analysis of proteins from a single hydrolysate. J. Biol. Chem. **251**:1936–1940.
20. **Szekely, E., and M. Simon.** 1983. DNA sequence adjacent to flagellar genes and evolution of flagellar-phase variation. J. Bacteriol. **155**:74–81.
21. **Wei, L. N., and T. M. Joys.** 1985. Covalent structure of three phase-1 flagellar filament proteins of *Salmonella*. J. Mol. Biol. **186**:791–803.
22. **Yanish-Perron, C., J. Vieira, and J. Messing.** 1985. Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors. Gene **33**:103–119.
23. **Yokoyama, S., A. Oobayashi, O. Tanabe, and E. Ichishima.** 1975. Action of crystalline acid carboxypeptidase from *Penicillium janthinellum*. Biochim. Biophys. Acta **397**:443–448.