29. E. M. Hedgecock, personal communication.
30. P. Homer and M. Chalfie, unpublished data.
31. There are only two *lin-32* mutations. The more severe of these affects all touch cells, but does not eliminate the anterior touch response ( *17*). Since the touch cells arise from similar, but not identical, lineages (*3, 4*), it is possible that the *lin-32* mutations reveal a lineage component necessary for the differentiation of the various touch cells. Alternatively, the posterior touch insensitivity of these mutants may result from a partial loss of gene activity. The *egl-5* mutations affect a number of cells in the tail (*16*). The PLM touch cells in these mutants appear normal, that is, they have both the large diameter microtubules and the mantle. Thus, we cannot determine whether the effect on touch sensitivity is direct (for example, causing an alteration in touch cell synapses) or indirect (for example, causing a displacement of the interneuronal processes onto which the touch cells must synapse).
32. M. Chalfie, J. N. Thomson, J. E. Sulston, *Science* **221**, 61 (1983).
33. J. G. White, E. Southgate, J. N. Thomson, S. Brenner, *Cold Spring Harbor Symp. Quant. Biol.* **48**, 633 (1983).
34. C. Q. Doe *et al.*, *Science* **239**, 170 (1988); C. Q. Doe, D. Smouse, C. S. Goodman, *Nature* **333**, 376 (1988); K. Blocklinger *et al.*, *ibid.*, p. 629.
35. M. Finney, G. Ruvkun, H. R. Horvitz, *Cell*, in press.
36. M. Ruiz-Gomez and J. Modolell, *Genes Dev.* **1**, 1238 (1987); R. Villares and C. V. Cabrera, *Cell* **50**, 415 (1987); R. L. Davis *et al.*, *ibid.* **51**, 987 (1987).
37. K. A. Wharton, K. M. Johansen, T. Xu, S. Artavanis-Tsakonas, *Cell* **43**, (1985); E. Hafen, K. Basler, J.-E. Edstroem, G. M. Rubin, *Science* **236**, 55 (19 I. S. Greenwald, *Cell* **43**, 583 (1985).
38. A. Tomlinson and D. F. Ready, *Science* **231**, 400 (1986).
39. L. W. Nawrocki, E. Southgate, J. N. Thomson, personal communication.
40. Reviewed in P. W. Ingham, *Nature* **335**, 25 (1988).
41. D. C. Sigurdson, G. J. Spanier, R. K. Herman, *Genetics* **108**, 331 (1984); E Rosenbluth, T. M. Rogalski, R. C. Johnsen, L. M. Addison, D. L. Baillie, G *Res.*, in press; T. M. Rogalski and D. L. Riddle, *Genetics* **118**, 61 (1988).
42. Genes have been positioned based on published data [(*2*); (M. L. Edgley and I Riddle, *Genet. Maps* **4**, 351 (1987)] as well as our own unpublished data depo with the Caenorhabditis Genetics Center (University of Missouri).
43. We are indebted to E. Bergholz, K. Buck, N. Hom, C. Masuoka, anc McDonald for technical assistance; to P. Brickman, P. Josephson, R. Goldstei Mindich, S. Shaham, and J. Srinivasan for the isolation and characterization o TR679-derived mutants; to M. Driscoll, E. Ferguson, L. Fischer, C. Savage Walthall, and J. Way for mapping data and helpful discussions; and to M. Le and S. Mount for suggestions on the manuscript. We are also grateful tc Caenorhabditis Genetics Center (University of Missouri) and our fellow *C. el* researchers for providing both Mec mutants and mapping strains. Supporte U.S. Public Health Service grant GM30997 to M.C.

Research Articles

# Reverse Transcriptase in a Clinical Strain of *Escherichia coli*: Production of Branched RNA-Linked msDNA

BERT C. LAMPSON, JING SUN, MEI-YIN HSU, JORGE VALLEJO-RAMIREZ, SUMIKO INOUYE, MASAYORI INOUYE

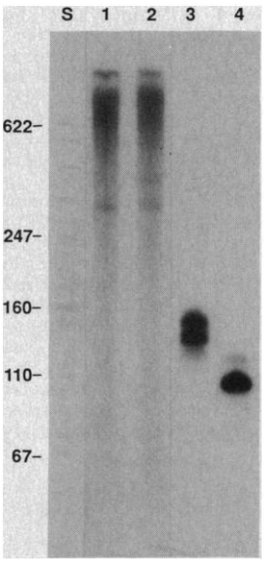
**Branched RNA-linked multicopy single-stranded DNA (msDNA) originally detected in myxobacteria has now been found in a clinical isolate of *Escherichia coli*. Although lacking homology in the primary structure, the *E. coli* msDNA is similar in secondary structure to the myxobacterial msDNA's, including the 2′,5′-phosphodiester linkage between RNA and DNA. A chromosomal DNA fragment responsible for the production of msDNA was cloned in an *E. coli* K12 strain; its DNA sequence revealed an open reading frame (ORF) of 586 amino acid residues. The ORF shows sequence similarity with retroviral reverse transcriptases and ribonuclease H. Disruption of the ORF blocked msDNA production, indicating that this gene is essential for msDNA synthesis.**

AN UNUSUAL SATELLITE DNA CALLED MSDNA (MULTICOPY SINGLE-STRANDED DNA) was originally found in *Myxococcus xanthus*, a Gram-negative bacterium living in soil (*1*). The satellite consists of a 162-base single-stranded DNA, the 5′ end of which is linked to a branched RNA (msdRNA) of 77 bases by a 2′,5′-phosphodiester linkage at the 2′ position of the 20th rG residue (*2*). There are approximately 700 copies of msDNA per genome. msDNA is widely distributed among various myxobact including the closely related *Stigmatella aurantiaca*, which has msDNA—msDNA-Sa163 (*3*), highly homologous to msDI Mx162 from *M. xanthus* (*4, 5*). Several *M. xanthus* strains, indep dently isolated from different sites, contain msDNA (*6*). We fo that *M. xanthus* contains another smaller species of msDNA form called mrDNA and now termed msDNA-Mx65 (*7*). In contras the close homology between msDNA-Mx162 and msDNA-Sa1 there is no primary sequence homology between msDNA-Mx and the small molecule, msDNA-Mx65. However, msDNA-M does share key secondary structures such as a branched rG residu DNA-RNA hybrid at the 3′ ends of the msDNA and the msdRI and stem-loop structures in RNA and DNA strands.

We have shown that msdRNA is derived from a much lor precursor RNA (pre-msdRNA), which can form a very stable st and-loop structure (*2*). A novel mechanism for msDNA synth was proposed, in which the stem-and-loop structure of msdRNA serves as a primer for initiating msDNA synthesis as as a template to form the branched RNA-linked msDNA, predicted that a reverse transcriptase (RT) is required for reaction (*2*). We now report that msDNA also exists in *Escher coli* and that a gene with sequence similarity to retroviral RT's

The authors are in the Department of Biochemistry, Robert Wood Johnson M School, University of Medicine and Dentistry of New Jersey, Piscataway, NJ 08

**Fig. 1.** Detection of msDNA in a clinical isolate of *E. coli*. Total RNA, prepared (*22*) from a 5-ml culture, was added to 50 μl of a reaction mixture containing: 50 m*M* tris-HCl (*p*H 8.3); 6 m*M* $MgCl_2$; 40 m*M* KCl; 5 m*M* DTT; 1 μ*M* dATP, dTTP, and dGTP; 0.04 μ*M* dCTP; 0.2 μ*M* [α-$^{32}$P]dCTP; and 10 units of AMV-RT (Boehringer Mannheim). The reaction mixture was incubated at 37°C for 30 minutes, then extracted with 50 μl of phenolchloroform (1:1), and precipitated with ethanol. The samples were subjected to electrophoresis on a 4 percent acrylamide–8*M* urea gel. (Lane S) Size markers in nucleotides; Msp I digest of pBR322 end-labeled with [α-$^{32}$P]dCTP and the Klenow fragment of DNA polymerase I; (lane 1) *E. coli* K12 strain C600; (lane 2) the same as in lane 1, except that the sample was treated with RNase A (5 μg, 10 minutes at 37°C) just before electrophoresis; (lane 3) clinical isolated Cl-1; and (lane 4) clinical isolate Cl-1 treated with RNase A. The clinical isolate was identified as *E. coli* (*9*).



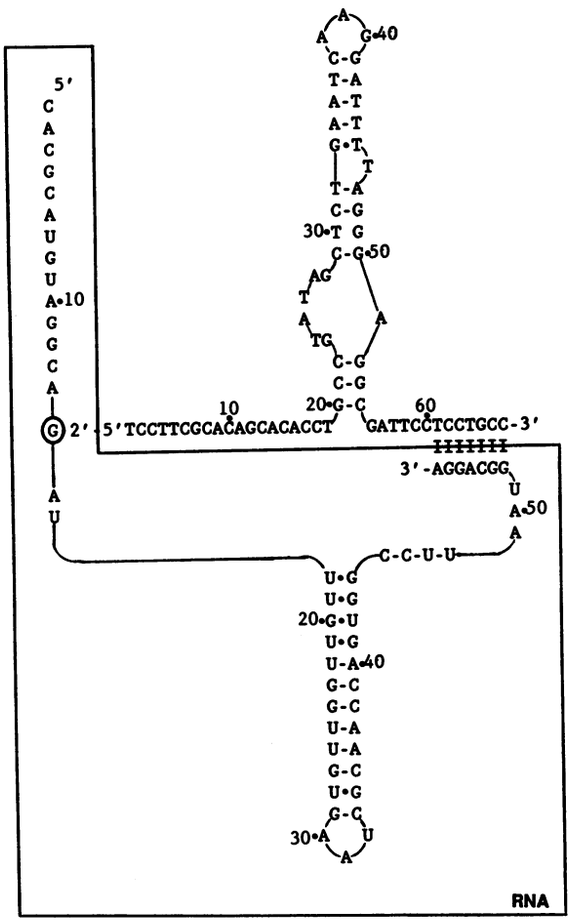
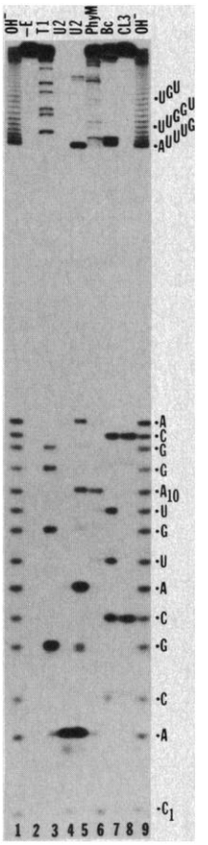




**Fig. 2.** The complete primary and proposed secondary structure of msDNA-Ec67. The DNA sequence was determined (*10*) with the use of 3′ -end labeled msDNA. The RNA sequence (msdRNA; boxed region) was determined with the use of base-specific ribonucleases as described (*2*). The 2′,5′ branched linkage between the 15th rG residue and the 5′ end of the DNA strand was determined with the debranching enzyme from HeLa cells as described previously (*2, 5, 11*). The branched rG at position 15 is circled, and both RNA and DNA are numbered from their 5′ ends. The secondary stem loop structure in both the DNA and RNA sequences were determined empirically on the basis of the already known structures of myxobacterial msDNA's.

**Fig. 3.** Determination of the RNA nucleotide sequence for the branched RNA linked to msDNA. Total RNA was prepared from the clinical strain Cl-1 and fractionated on a 5 percent acrylamide gel. msDNA containing full-length RNA was eluted from the gel. This fraction was then labeled at the 5′ end of the RNA with [γ-$^{32}$P]ATP and T4 polynucleotide kinase. The 5′ end-labeled RNA linked to msDNA was again purified on an 18 percent acrylamide–8*M* urea sequencing gel. The labeled RNA was then sequenced with limited digestion by base-specific RNases (*2*). (Lanes $OH^-$) Partial alkaline hydrolysis ladder (0.5*M* sodium bicarbonate/buffer *p*H 9.2); (lane E−) no enzyme treatment of the labeled RNA linked to msDNA; (lane T1) RNase T1 (1 U per reaction, 55°C, 15 minutes); (lane U2) RNase U2 (1 U and 0.5 U per reaction, 55°C, 15 minutes); (lane PhyM) RNase PhyM (1 U per reaction, 55°, 15 minutes); (lane Bc), RNase B. cerus (2 U per reaction, 55°C, 15 minutes); (lane CL3) RNase CL3 (2 U per reaction, 37°C, 15 minutes). The large gap in the sequence gel is due to msDNA linked at the rG residue at position 15 by a 2′,5′ phophodiester linkage (*5*). The RNA sequence at the 3′ end region from the branched rG residue (the upper part of the gel) was determined from a 6 percent gel.




ribonucleases H is required for the production of msDNA.

**msDNA in *E. coli*.** The recent serendipitous finding of msDNA (msDNA-Ec86) in *E. coli* B by Lim and Maas (*8*) prompted us to search for msDNA in other *E. coli* strains. As shown by Yee *et al.* (*1*), msDNA is not found in the common laboratory strain K12; however, we have now found msDNA in a clinical *E. coli* strain isolated from a patient with a urinary tract infection. Fifty independent *E. coli* urinary tract isolates were examined for the presence of msDNA (*9*). The screening method consisted of treatment of total RNA prepared from each strain with RT from avian myeloblastosis virus (AMV) in the presence of [α-$^{32}$P]dCTP (deoxycytidylate) plus the deoxynucleotides dATP, dTTP, and dGTP and subsequent polyacrylamide gel electrophoresis (PAGE). Since msDNA contains a DNA-RNA duplex structure, the 3′ end of the DNA molecule serves as an intramolecular primer and the RNA molecule serves as a template for RT. When RNA prepared from one of the clinical strains, *E. coli* Cl-1 (*9*), was labeled in this manner, two distinct, low molecular weight bands of about 160 bases became labeled with $^{32}$P (Fig. 1). If the labeled sample is digested with ribonuclease (RNase) A before it is placed on the gel, a single band corresponding to 105 bases of single-stranded DNA is detected (Fig. 1, lane 4). This indicates that both bands in lane 3 contain a single-stranded DNA of identical size. The two labeled bands observed before RNase treatment (lane 3) are due to two species of msDNA comprising a single species of single-stranded DNA linked to RNA molecules of two different sizes. RNA molecules of two different sizes have been observed at the 5′ ends of msDNA from myxobacteria in which a precursor molecule contains a longer RNA that is processed into a smaller mature form (*2, 5*). Among the 89 clinical isolates screened, three other strains produced msDNA-like molecules of varying size and quantity, suggesting extensive diversity among these molecules.

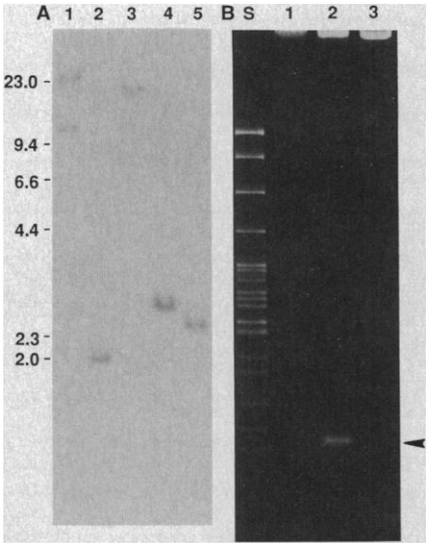



**Fig. 4.** DNA blot analysis of *E. coli* Cl-1 chromosomal DNA(A) and analysis of msDNA synthesis by pCl-1E and pCl-1P(B). (**A**) The chromosomal DNA was digested with Eco RI (lane 1), Hind III (lane 2), BamHI (lane 3), PstI (lane 4), and BglII (lane 5). For each lane, 3 μg of the DNA digest was applied to a 0.7 percent agarose gel. After electrophoresis, the gel was blotted to a nitrocellulose filter, and hybridization analysis was carried out (*23*) with msDNA labeled by AMV-RT and with [α-$^{32}$P]dCTP as a probe. Numbers at the left represent the molecular sizes in kilobases. (**B**) Total DNA prepared from each strain was treated with RNase A, separated on a 5 percent acrylamide gel, and stained with ethidium bromide (lane S) pBR322 digested with Msp I size markers in base pairs; (lane 1) DNA prepared from the host strain CL-83 (*recA*); (lane 2) CL-83 (*recA*) transformed with plasmid pCl-1E (11.6-kb Eco RI fragment as in Fig. 5); (lane 3) with plasmid pCl-1P (2.8-kb Pst I (a)–Pst I (b) fragment as in Fig. 5). An arrow indicates the position of msDNA.

As reported earlier (*6*), msDNA was not observed in the *E. coli* K-12 strain, C600 (Fig. 1, lanes 1 and 2).

**Nucleotide sequence of msDNA Ec-67.** To determine the base sequence of the DNA molecule, we isolated the RNA-DNA complex from the clinical strain and labeled it at the 3′ end of the DNA molecule with AMV-RT and [α-$^{32}$P]dATP. Supplementing the reaction mixture with dideoxy-CTP (ddCTP), ddTTP, and ddGTP resulted in the addition of a single-labeled adenine at the 3′ end of the DNA strand. The RNA was removed with RNase A plus RNase T1 and the end-labeled DNA was sequenced (*10*). The msDNA consists of a single-stranded DNA of 67 bases and, as in the case of msDNA's from myxobacteria (*1, 2*), it can form a secondary hairpin structure (Fig. 2). The primary sequence, however, is not homologous to any of the myxobacterial msDNA's, nor to the msDNA from *E. coli* B (msDNA-Ec86) (*8*).

The sequence of the RNA molecule was determined with the RNA-DNA complex purified from *E. coli* Cl-1. The RNA sequence was determined with base-specific RNases as described (*7*). A large gap is observed in the RNA sequence ladder (Fig. 3); this gap is due to the DNA strand branched at the 2′ position of the 15th rG residue of the RNA strand which produces a shift in mobility of the sequence ladder (Fig. 2). The RNA consists of 58 bases with the DNA molecule branched at the G residue at position 15 by a 2′,5′-phosphodiester linkage. The branched G structure was determined as previously described for msDNA's from myxobacteria (*2, 5*). After treatment with RNases A and T1, the msDNA retained a small oligoribonucleotide linked to the 5′ end of the DNA molecule because of the inability of RNases to cleave in the vicinity of the branched linkage. The 5′ end was labeled with [γ-$^{32}$P]ATP, in the presence of T4 polynucleotide kinase, and the labeled RNA molecule was detached from the DNA strand by a debranching enzyme purified from HeLa cells (*11*). This small RNA was a tetraribonucleotide that could be digested with RNase T1 to yield a labeled dinucleotide. Since RNase T1 could not cleave the RNA molecule at the G residue before the debranching enzyme treatment, it was concluded that the single-stranded DNA is branched at the G residue via a 2′,5′-phosphodiester linkage. In addition, partial RNase $U_2$ digestion cleaved the RNA molecule to yield a $^{32}$P-labeled mononucleotide and a $^{32}$P-labeled trinucleotide. Thus, the sequence of the tetranucleotide is $^{5'}$A-G-A-(U or C)$^{3'}$ (Fig. 2). Despite a lack of primary structural homology, msDNA-Ec67 displays all the distinctive features found in msDNA's from myxobacteria. These include a single-stranded DNA with a stem-and-loop structure, a single-stranded RNA with a stem-and-loop structure, a 2′-5′-phosphodiester linkage between the RNA and DNA, and a DNA-RNA hybrid at their 3′ ends. This hybrid structure was confirmed by demonstrating sensitivity of the RNA molecule to RNase H.

**Cloning of the locus for msDNA-Ec67.** In order to identify the DNA fragment that is responsible for msDNA synthesis in *E. coli* Cl-1, we performed DNA blot hybridizations with various restriction enzyme digests of total chromosomal DNA prepared from *E. coli* Cl-1; we used msDNA-Ec67 labeled with AMV-RT (the same preparation as shown in Fig. 1, lane 3) as a probe (Fig. 4A). Digestion with Eco RI (lane 1), Hind III (lane 2), Bam HI (lane 3), Pst I (lane 4), and Bgl II (lane 5) showed single band hybridization signals corresponding to 11.6, 2.0, ~22, 2.8, and 2.5 kilobase pairs (kb), respectively. The upper band appearing in the Eco RI digestion is due to incomplete digestion of the chromosomal DNA. Analysis of total chromosomal DNA prepared from *E. coli* Cl-1 by agarose gel electrophoresis revealed that the strain contains two plasmids of different size. However, neither plasmid hybridized with the $^{32}$P-labeled probe, an indication that the fragments detected in Fig. 4A are derived from chromosomal DNA. Furthermore, there is only one location for the msDNA-coding region on the chromosome, since various restriction enzyme digestions gave only one band of varying sizes. Similar results were observed for the msDNA's of myxobacteria (*1, 5, 7*).

The 11.6-kb Eco RI fragment and the 2.8-kb Pst I fragment were each cloned into pUC9 (*12*), and *E. coli* CL83, which is a *recA* transductant of strain JM83 and an msDNA-free K-12 strain (Fig.

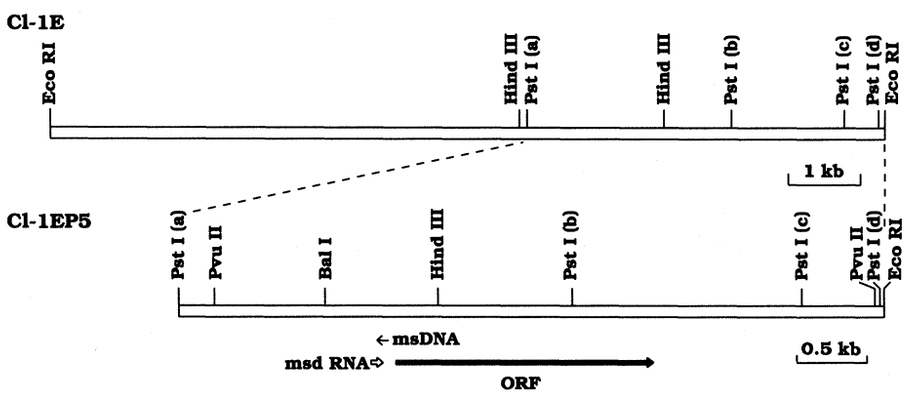



**Fig. 5.** Restriction map of the 11.6-kb Eco RI fragment. In the Cl-1E map, the left-hand half (Eco RI to Hind III) was not mapped. In the Cl-1EP5 map, the locations and the orientations of msDNA and msdRNA are indicated by a small arrow and an open arrow, respectively. A large solid arrow represents an ORF and its orientation.

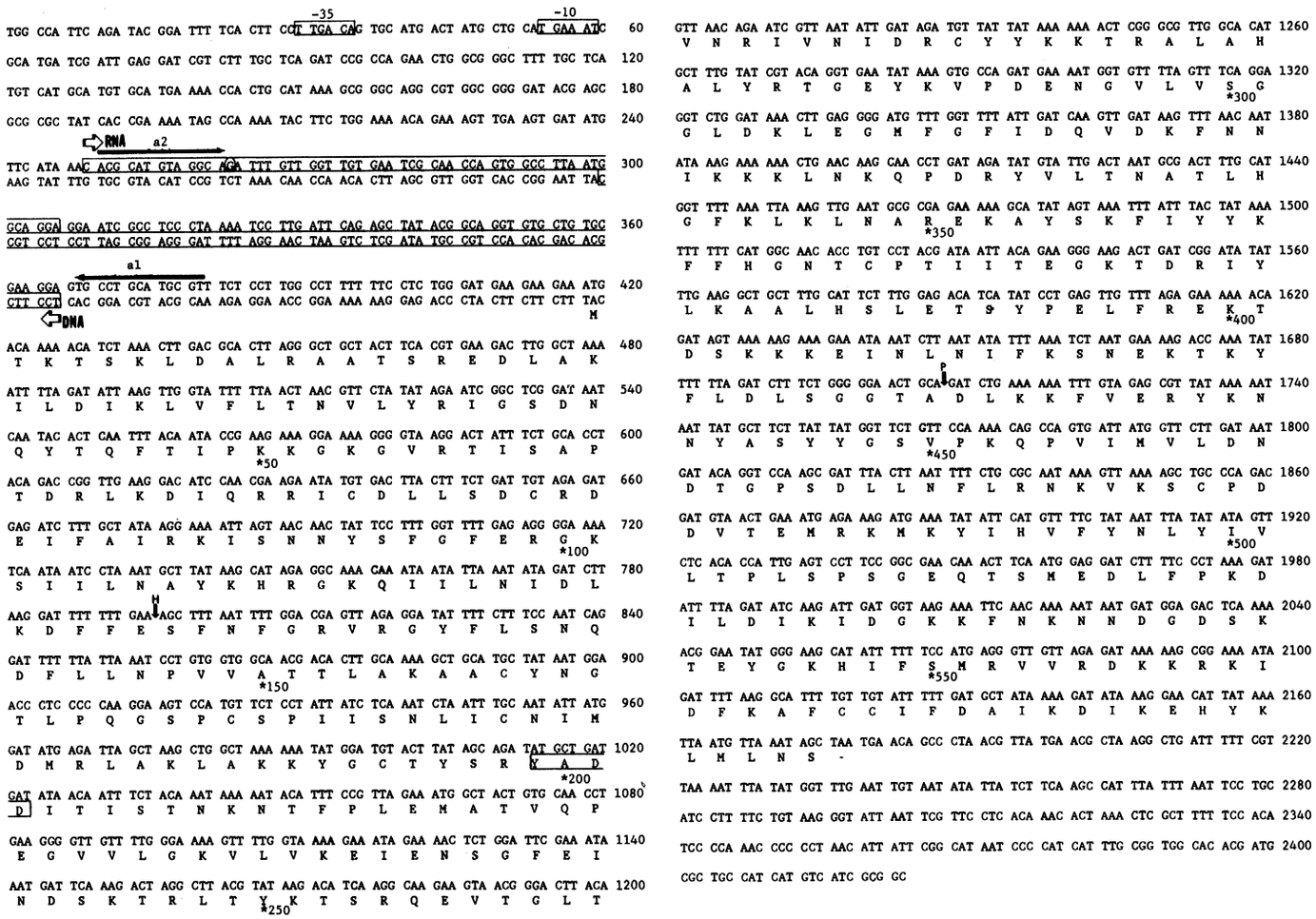


**Fig. 6.** Nucleotide sequence of the region from the *E. coli* Cl-1 chromosome encompassing the msDNA and msdRNA coding regions and an ORF downstream of the msdRNA region. The entire upper strand beginning at the Bal I site (Fig. 5) and ending just beyond the ORF is shown. Only a part of the complementary lower strand is shown from base 241 to 420. The long boxed region of the upper strand (249 to 306) corresponds to the sequence of the branched RNA (msdRNA) (Fig. 2). The boxed region of the lower strand corresponds to the sequence of the DNA portion of msDNA (Fig. 2). The starting site for DNA and RNA and the 5′ to 3′ orientations are indicated by large open arrows. The msdRNA and msDNA regions overlap at their 3′ ends by seven bases. The circled G residue at position 263 represents the branched rG of RNA linked to the 5′ end of the DNA strand in msDNA. Long solid arrows labeled a1 and a2 represent inverted repeat sequences that may be important in the secondary structure of the primary RNA transcript involved in the synthesis of msDNA (*2*). The nucleotide at position 257 (U on the RNA transcript) and that at position 373 (G on the RNA transcript) form an U · G pair in the stem between sequence a1 and a2. The proposed promoter elements (−10 and −35 regions) for the primary RNA transcript are also boxed. The ORF begins with the initiation codon at base 418. The YXDD amino acid sequence conserved among known RT proteins is boxed. Numbers on the right signify the nucleotide bases and those with an asterisk represent amino acids. Small vertical arrows labeled H and P locate the Hind III and Pst I restriction cleavage sites, respectively. The DNA sequence was determined by the chain termination method (*24*) with the use of synthetic oligonucleotides as primers. Abbreviations for the amino acid residues are: A, Ala; C, Cys; D, Asp; E. Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L. Leu; M, Met; N, Asn; P, Pro; Q, Gln; R. Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.

4B, lane 1), was transformed with the plasmids. Cells transformed with the 11.6-kb Eco RI clone (pCl-1E) produced msDNA (Fig. 4B, lane 2), whereas cells transformed with the 2.8-kb Pst I clone (pCl-1P) failed to produce any detectable msDNA (Fig. 4B, lane 3). The 11.6-kb fragment was analyzed (Fig. 5). A DNA blot analysis of the fragment revealed that a 1.8-kb Pst I–Hind III fragment hybridized with the msDNA probe. When the DNA sequence of this fragment was determined, a region identical to the sequence of the msDNA molecule was discovered. The DNA sequence corresponding to the sequence of msDNA is indicated by the enclosed box on the lower strand in Fig. 6 and the orientation is from right to left. The location of this sequence is also indicated by a small arrow in Fig. 5. As is the case for all other known myxobacterial msDNA's (*2, 5, 7*), a sequence identical to that of the RNA linked to msDNA (Fig. 2) was found downstream of the msDNA-coding region in opposite orientation and overlapping with that region by seven bases. This sequence is indicated by the enclosed box on the upper strand in Fig. 6 and the branched G residue is circled. Again, as in all the msDNA's in myxobacteria, there is an inverted repeat comprised of a 13-base sequence immediately upstream of the branched G residue (residue 250 to 262; sequence a2 in Fig. 6) and a sequence at the 3′ end shown by an arrow in Fig. 6 (residue 368 to 380; sequence a1). As a result of this inverted repeat, a putative longer primary RNA transcript beginning upstream of the RNA coding region and extending through the msDNA coding region would be able to self-anneal and form a stable secondary structure, which is proposed to serve as the primer as well as the template for msDNA synthesis (*2*).

**Existence of an essential gene for msDNA synthesis.** The proposed model for the synthesis of msDNA (*2*) predicts that an RT plays an essential role in the production of msDNA. Because of similarity in the structural arrangement of the msDNA coding regions between myxobacteria and *E. coli*, it is reasonable to assume that an RT is also required for the synthesis of msDNA-Ec67.

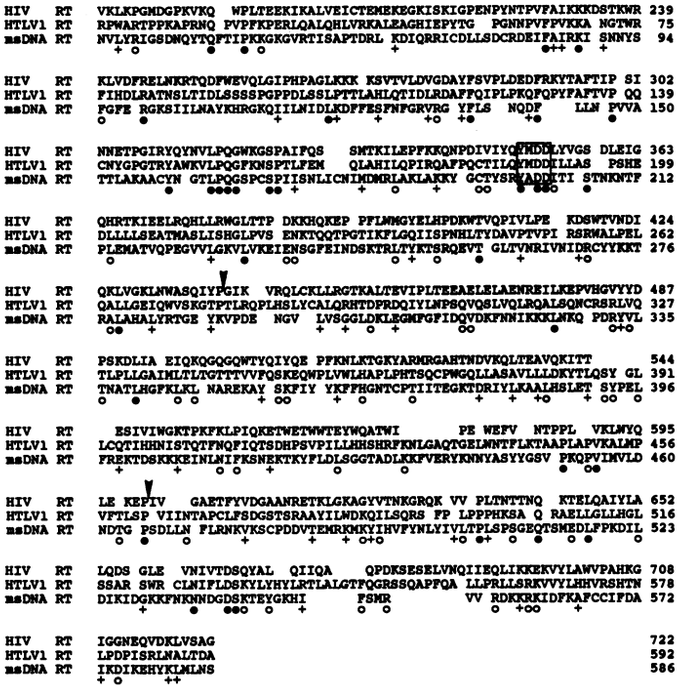


**Fig. 7.** Amino acid sequence alignment of the *E. coli* msDNA ORF with a portion of the retroviral Pol sequence from HIV and HTLV-I. Amino acid sequences are compared with matching residues assigned as follows: (+) amino acid common to msDNA and HIV RT's; (○) amino acid shared by msDNA and HTLV-I RT's; and (●) amino acid shared by all three proteins. Arrows divide the protein sequences into three functional domains (*15, 16*): an amino terminal RT domain, a carboxyl-terminal RNase H region, and a central "tether" region. The specific amino acid residues for the RT, tether, and RNase H domains, for each protein are: HIV, 177 to 439, 440 to 600, and 601 to 722, respectively; HTLV-I, 15 to 277, 278 to 462, and 463 to 592, respectively; msDNA ORF, 32 to 290, 291 to 465, 466 to 586, respectively. The YXDD polymerase consensus sequence is outlined with a box.

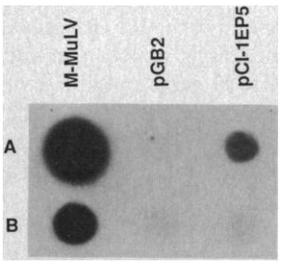



**Fig. 8.** Detection of RT activity from various cell extracts. Crude cell extracts were prepared from *E. coli* strain C2110 (*polA*$^-$) (*17*) containing plasmid pCl-1EP5 encoding the msDNA-ORF (Fig. 5) as well as the vector plasmid (pGB2) alone. Extracts were also prepared from the *E. coli* strain PRTS7-1 (*polA*+) containing the cloned M-MuLV RT gene (*16, 17*). Crude extracts were prepared essentially as described (*25*). Crude extract equivalent to 15 μg of total protein was added to a 50-μl reaction cocktail (50 m*M* tris-HCl, *p*H 7.8, 10 m*M* DTT, 60 m*M* NaCl, 0.05 percent NP-40, 10 m*M* $MgCl_2$, 0.5 μg of poly(rC) · oligo(dG), and 0.1 μ*M* [α-$^{32}$P]dGTP) and incubated at 37°C for 1 hour. A sample (5 μl) of the reaction mixture was then spotted onto DEAE paper (DE81, Whatman). The paper was washed to remove unincorporated label (*17*) and then exposed to an x-ray film. (**A**) All reactions contain added template-primer poly(rC) · oligo(dG). (**B**) Control reactions in which no template-primer is added. Columns contain the designated cell extracts: M-MuLV, cloned Moloney murine leukemia virus RT gene; pGB2 (*26*), vector plasmid in strain C2110; pCL-1EP5, recombinant plasmid with the cloned msDNA gene. The high background activity observed with the M-MuLV control extract can be attributed to the activity of DNA polymerase I since this extract is obtained from a PolA$^+$ strain (HB101).

The 2.8-kb Pst I fragment [from Pst I (a) to Pst I (b) in Fig. 5] was not able to synthesize msDNA. However, an overlapping 3.9-kb fragment from Bal I [1.0-kb downstream of Pst I (a) in Fig. 5] to the following Eco RI site contains all the information required for synthesis of msDNA. This indicates that a region downstream of the Pst I site (Fig. 5) is required for msDNA production. The nucleotide base sequence from this region revealed a long open reading frame (ORF) of 586 amino acid residues, starting with the initiation codon ATG at nucleotide 418 to 420 (Fig. 6). A distance of only 51 bases separates the initiation codon from the region that encodes msDNA. A putative Shine-Dalgarno sequence (GGA) can be found ten bases upstream of the initiation codon. When the *lacZ* gene was fused in frame at the Hind III site (within the ORF) at amino acid residue 126, β-galactosidase activity was detected. Thus the region encompassing the ORF is indeed transcribed and the gene product encoded by the ORF is essential for msDNA synthesis. In a preliminary experiment, both msdRNA and the ORF appeared to be transcribed as the same transcription unit, since a deletion mutation removing the sequence from residue 1 to 181 blocked the expression of the *lacZ* gene fused at the Hind III site. A putative promoter can be found in the deleted sequence (Fig. 6). These −35 and −10 regions probably serve as the promoter for both msdRNA synthesis and the ORF.

**Sequence similarity with retroviral reverse transcriptases.** When the amino acid sequence of the ORF was compared with known proteins, a striking similarity was found with retroviral RT's. Comparison of the ORF with RT's from HIV (human immunodeficiency virus) (*13*), and HTLV-I (human T cell leukemia virus type I) (*14*) shows that the first domain (Asn$^{32}$ to Val$^{291}$) of the ORF matches well with the RT domains of HIV and HTLV-I (Fig. 7). In particular, the sequences around the polymerase consensus Asp-Asp sequence (*15*) (boxed in Figs. 6 and 7) are well conserved. Of 260 amino acid residues in this domain, 44 and 38 residues are identical with HIV and HTLV-I, respectively. Between the RT's of these retroviruses, there are 78 identical amino acid residues in this domain.

The *pol* gene of retroviruses produces a protein consisting of RT and RNase H activities; the former at the $NH_2$-terminal and the latter at the COOH-terminal region of the *pol* gene product (*13, 16*). These domains are separated by a poorly conserved "tether" domain of approximately 160 to 190 amino acid residues (*13*). On the basis of the HIV sequence, the similarities (only identical amino acid residues) between HIV and HTLV-I are 29.5 and 16.8 percent for the RT domain and the tether domain, respectively. The similarities between HIV and msDNA are 16.9 and 10.3 percent for the RT domain and the tether domain, respectively. The similarities between HTLV-I and msDNA are 14.6 and 15.5 percent for the RT domain and the tether domain, respectively. These results indicate that, in addition to the RT region, there are reasonable similarities in the tether domain between retroviruses and msDNA. An alignment of the RNase H domains also revealed that there are similarities between retroviruses and msDNA (15.7 and 17.4 percent with HIV and HLTV, respectively) (Fig. 7). The similarity between HIV and HTLV-I in this region is 18.0 percent.

Cell extracts were prepared and assayed for RT activity associated with the production of msDNA as predicted from the amino acid homologies. Only the *E. coli* strain (C2110, *polA*) (*17*) harboring the plasmid pCl-1EP5 containing the msDNA ORF displayed RT activity (Fig. 8). In the RT assay, the *polA* strain was used to eliminate high background activity due to DNA polymerase I. No RT activity was detected in extracts containing the vector plasmid alone, or when the template-primer [poly(rC-dG)] was absent from the reaction mix (Fig. 8). The Pst I site (b in Fig. 5) is located at amino acid residue 430, which is between the tether domain and the RNase H domain. A plasmid lacking sequences downstream of this Pst I site did not produce msDNA. Accordingly, the RNase H domain may be essential for msDNA synthesis or, alternatively, Pst I disruption may result in inactivation of RT.

In addition to the similarity between msDNA-Ec67 RT and

retroviral RT, there is a similarity between msDNA and retroviruses; DNA synthesis starts at the site upstream of the RT–RNase H gene, and the orientation of DNA synthesis is opposite to the direction of transcription of the RT–RNase H gene. In the case of retroviruses, transfer RNA's are recruited from the cell for the priming reaction (*18*), whereas for msDNA an RNA transcript serving as template also serves as a primer by self-annealing to form a stable secondary structure (*2, 5*).

**Origin of the *E. coli* reverse transcriptase.** At present the relation between msDNA and retroviruses is not known. Nor is it known why some of the clinical *E. coli* strains isolated from human patients produce msDNA. Our data indicate that msDNA's produced by four independent *E. coli* strains, all isolated from urinary infections, share little homology. This suggests that there may be large numbers of species of msDNA in *E. coli*. In contrast to msDNA's found in *E. coli*, msDNA-Mx162 from *M. xanthus* is highly conserved, since nine independent *M. xanthus* strains isolated from various sites have msDNA that hybridizes with the original msDNA-Mx162 (*6*). Furthermore, msDNA from another myxobacterium, *S. aurantiaca* (msDNA-Sa163) (*5*) also shows a high degree of homology to msDNA-Mx162 (*4*).

Several lines of evidence suggest that the RT gene found in the *E. coli* strain C1-1 is not likely to have originated in *E. coli*, but rather was recently acquired from some other source. For example, only about 4 percent of *E. coli* strains tested were found to produce msDNA. In addition, the RT gene from strain C1-1 does not cross-hybridize to chromosomal DNA from four other *E. coli* strains that produce msDNA molecules, an indication that there is extensive diversity among these RT genes. In contrast, a DNA fragment from the *E. coli* K12 sigma factor gene can hybridize to chromosomal DNA from all five msDNA-producing *E. coli* strains, indicating the conserved nature of sigma factors. An analysis of the *E. coli* RT gene indicates that the codon usage for this gene is quite different from most *E. coli* proteins. In particular, AGA and AGG, the least frequently (2.7 percent) used codons for arginine among 199 *E. coli* genes (*19*), occurs at a frequency of 64.5 percent in the *E. coli* RT gene. Similarly, CUG is the most commonly used codon for leucine (61.3 percent) (*19*) in *E. coli* genes, while its prevalence in the RT gene is only 9.1 percent. The AT base pair content of the *E. coli* RT gene was calculated to be 67.6 percent, which is substantially higher than the AT content of the *E. coli* genome (49 percent) (*20*). The AT contents of HIV and HTLV-I RT genes are 62.1 and 47.8 percent, respectively.

Many questions remain to be answered, including (i) are there any particles associated with msDNA; (ii) is the msDNA region transposable like the Ty element of yeast (*21*); (iii) can the element responsible for the production of msDNA be transferred from cell to cell; (iv) can a RT from one strain (*E. coli* or myxobacteria) complement the production of msDNA of other strains; (v) does the promoter for the RNA transcript have any similarities to the retroviral LTR; (vi) are there any specific integration sites for the msDNA element on the *E. coli* chromosome; (vii) why is the branched G residue conserved; (viii) is there an enzyme responsible for priming DNA synthesis at the 2′-OH position of the rG residue; (ix) why and how does msDNA synthesis stop at one distinct site on the RNA template; and (x) how different biochemically are the msDNA RT's from retroviral RT's?

The existence of RT in prokaryotes, previously predicted (*2*), is now evident from our data. This fact raises intriguing questions concerning possible roles of this enzyme in the prokaryotes other than msDNA production. We have also found that *M. xanthus* in which msDNA was originally discovered has a long ORF in the same manner as found for msDNA-Ec67 (*27*). This ORF has a high degree of similarity to the *E. coli* RT. Since eight independent isolates of *M. xanthus* produce homologous msDNA, the *M. xanthus* RT is likely to have been acquired at a very early stage of its evolution in contrast to the *E. coli* RT.

## REFERENCES AND NOTES


1. T. Yee, T. Furuichi, S. Inouye, M. Inouye, *Cell* **38**, 203 (1984).
2. A. Dhundale *et al.*, *ibid.* **51**, 1105 (1987).
3. Because of the growing number of newly discovered species of msDNA, a new designation for this molecule has been adopted based on the host organism in which it is found and the size of its single-stranded DNA. For example, msDNA from *M. xanthus* is now msDNA-Mx162.
4. T. Furuichi, A. Dhundale, M. Inouye, S. Inouye, *Cell* **48**, 47 (1987).
5. T. Furuichi, M. Inouye, S. Inouye, *ibid.*, p. 55.
6. A. Dhundale, T. Furuichi, S. Inouye, M. Inouye, *J. Bacteriol.* **164**, 914 (1985).
7. A. Dhundale, M. Inouye, S. Inouye, *J. Biol. Chem.* **263**, 9055 (1988).
8. D. Lim and W. Maas, *Cell*, in press.
9. The clinical *E. coli* strains were urinary tract isolates provided by M. Weinstein from the microbiology laboratory, R. W. Johnson Hospital, New Brunswick, NJ. The clinical strain Cl-1 was identified with the use of the API-20E identification system (API Laboratory Products) and gave a typical *E. coli* profile number of 5044552.
10. A. M. Maxam and W. Gilbert, *Methods Enzymol.* **65**, 499 (1980).
11. B. Ruskin and M. Green, *Science* **229**, 135 (1985); J. Arenas and J. Hurwitz, *J. Biol. Chem.* **262**, 4274 (1987); The debranching enzyme was a gift from J. Hurwitz.
12. C. Yanisch-Perron, J. Vieria, J. Messing, *Gene* **33**, 103 (1985).
13. L. Ratner *et al.*, *Nature* **313**, 277 (1985); M. S. Johnson, M. A. McClure, D.-F. Feng, J. Gray, R. F. Doolittle, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 7648 (1986).
14. M. Seiki, S. Hattori, Y. Hirayama, M. Yoshida, *Proc. Natl. Acad. Sci. U.S.A.* **80**, 3618 (1983); R. Patarca and W. A. Haseltine, *Nature* **309**, 728 (1984).
15. H. Toh, H. Hayashida, T. Miyata, *Nature* **305**, 827 (1983); D.-F. Geng, M. S. Johnson, R. F. Doolittle, *J. Mol. Evol.* **21**, 113 (1985).
16. H. E. Varmus, *Nature* **314**, 583 (1985); N. Tanese and S. P. Goff, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 1777 (1988).
17. N. Tanese, M. Roth, S. P. Goff, *Proc. Natl. Acad. Sci. U.S.A.* **82**, 4944 (1985); N. Tanese, J. Sodroski, W. A. Haseltine, S. P. Goff, *J. Virol.* **59**, 743 (1986); *E. coli* strain C2110 (polA1$^-$) was provided by M. Roth and S. Goff.
18. N. Weiss, H. Teich, J. Varmus, J. Coffin, Eds., *RNA Tumor Viruses*, vol. 2, Supplements and Appendixes (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, ed. 2, 1985).
19. T. Maruyama, T. Gojobori, S. Aota, T. Ikemura, *Nucleic Acids Res.* **14** (suppl.), r151 (1986).
20. G. Fasman, Ed., *CRC Handbook of Biochemistry and Molecular Biology, Nucleic Acids* (CRC Press, Cleveland, 1976), vol. 2, p. 102.
21. J. D. Boeke, D. J. Gorfinkel, C. A. Styles, G. R. Fink, *Cell* **40**, 491 (1985); D. J. Eichinger and J. D. Boeke, *ibid.* **54**, 955 (1988).
22. T. Maniatis, E. F. Fritsch, J. Sambrook, in *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 1982).
23. E. Southern, *J. Mol. Biol.* **98**, 503 (1975).
24. F. S. Sanger, S. Nicklen, A. R. Coulson, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463 (1977).
25. M. J. Roth, N. Tanese, S. P. Goff, *J. Biol. Chem.* **260**, 9326 (1985); A. Hizi, C. McGill, S. H. Hughes, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 1218 (1988).
26. G. Churchward, D. Belin, Y. Nagamine, *Gene* **31**, 165 (1984).
27. S. Inouye, M-Y. Hsu, S. Eagle, M. Inouye, *Cell*, in press; B. Lampson, M. Inouye, S. Inouye, *ibid.*, in press.


28. We thank A. Shatkin, D. Reinberg, and C. Lerner, for critical reading of this manuscript; S. Eagle for skillful assistance; W. Maas and D. Lim for sharing their findings prior to publication, which made our work possible; M. Roth for *E. coli* strain C2110 and helpful discussions; S. Goff for the M-MuLV clone, strain PRTs7-1; and E. Arnold for his assistance in the computer alignment of amino acid sequences. Supported by USPHS grant GM26843, a grant from Takara Shuzo Co., Ltd., and PHS research fellowship grant F32 GM11970-01A1.

27 December 1988; accepted 28 January 1989