The EMBO Journal Vol.2 No.6 pp.967-971, 1983

Nucleotide sequence of the structural gene for dUTPase of *Escherichia coli* K-12

Lennart G. Lundberg*, Hans-Olof Thoresson, Olle H. Karlström[1] and Per Olof Nyman[2]

Department of Biochemistry and Biophysics, Chalmers Institute of Technology, S-412 96 Göteborg, Sweden, [1]University Institute of Microbiology, Øster Farimagsgade 2A, DK-1353 Copenhagen, Denmark, and [2]Department of Biochemistry, Chemical Center, University of Lund, PO Box 740, S-220 07 Lund, Sweden

Communicated by Bo G. Malmström
Received on 15 March 1983; revised on 14 April 1983

The nucleotide sequence of the dUTPase structural gene, *dut*, of *Escherichia coli* has been determined. The DNA sequence predicts a polypeptide chain of 150 amino acid residues (mol. wt. 16 006) corresponding in size and composition to the purified dUTPase subunit. In a tentative promoter region preceding the *dut* gene, the −35 and −10 regions are separated by a *Sac*I (*Sst*I) site. Cloning of the *dut* gene utilizing this *Sac*I site was previously shown to reduce *dut* expression dramatically. The nucleotide sequence also contains a 210-codon open reading frame 106 bp downstream of *dut* and co-directional with *dut*. Previous protein synthesis experiments using *dut* plasmids allocated the gene of a polypeptide of mol. wt. 23 500 to this DNA region. The open reading frame thus may correspond to a protein of unknown function co-transcribed with the *dut* gene.

Key words: DNA sequence/*dut*/dUTP/dUTPase/restriction endonuclease mapping

Introduction

Deoxyuridine 5′-triphosphate nucleotidohydrolase (dUTPase, EC 3.6.1.23) is an enzyme involved in nucleotide metabolism (Bertani *et al.*, 1961; Shlomai and Kornberg, 1978). It has two main functions: (i) it produces dUMP, the immediate precursor of thymidine nucleotides and (ii) it decreases the intracellular concentration of dUTP.

The structural gene for dUTPase, *dut*, maps at 81 minutes on the *Escherichia coli* linkage map (Bachmann *et al.*, 1980) close to the *pyr*E gene (Tye *et al.*, 1977). The gene has been cloned on different plasmids (Lundberg *et al.*, 1979; Taylor *et al.*, 1980); Lundberg *et al.*, 1983a), and extensive restriction endonuclease mapping has been carried out (Lundberg *et al.*, 1983b). dUTPase has been purified to homogeneity and its amino acid composition determined (Shlomai and Kornberg, 1978). Plasmids allowing efficient overproduction should provide an important tool for further enzymological studies of dUTPase.

We have determined the nucleotide sequence of the *dut* region to allow access to the amino acid sequence of dUTPase and to decide if there is an unknown gene, co-transcribed with *dut*, and positioned between *dut* and *pyrE* as previously suggested (Lundberg *et al.*, 1983b).

*To whom reprint requests should be sent.

Table I. *E. coli* K-12 strains and plasmids

Designation	Bacterium	Plasmid	Relevant genotype	Source or reference
KK975	NF929	pKK6	*Tc*^{r}, *dut*$^+$, *pyrE*$^+$	Lundberg *et al.*, 1979
JF1600	NF930	pGA2	*Tc*^{r}, *pyrE*$^+$	An *et al.*, 1979
KK2200	NF929	pLT1	*Ap*^{r}, *dut*$^+$	This study
NF929	NF929	–	*thr leu his argH pyrE thi*	Fiil *et al.*, 1977
NF930	NF930	–	NF929 *spoT*	An *et al.*, 1979

Results

Isolation of the dut gene region

Our starting material for the nucleotide sequencing of the *dut* gene was three different plasmids carrying the *dut* gene or part of it (Figure 1). The map of plasmid pKK6 extends and corrects the map presented earlier (Lundberg *et al.*, 1983b). From these plasmids the following DNA fragments were prepared and later used for subcutting with different restriction endonucleases and labeling: from pKK6 a 1.8-kb *Pvu*II-*Pvu*II fragment, a 1.0-kb *Sac*I-*Pvu*II fragment and a 0.66-kb *Pvu*II-*Bam*HI fragment, from pLT1 a 1.2-kb *Eco*RI-*Pvu*II fragment and from pGA2 a 1.6-kb *Bam*HI-*Bam*HI fragment.

Nucleotide sequence

Since the *Sac*I site in the *dut* region had been found to be important for the expression of dUTPase activity (Lundberg *et al.*, 1983a) we concentrated on sequencing the DNA region around the *Sac*I site and to the right of it (Figure 2). Figure 3 presents the nucleotide sequence and the amino acid sequence of dUTPase which can be derived from the nucleotide sequence. The numbering of the DNA sequence starts from an *Asu*I site (at the last nucleotide in the recognition sequence) used for labeling and extends 1609 bp (towards the *pyrE* gene).

The nucleotide sequence contains two open reading frames downstream from the *Sac*I site. One starts with an ATG at position 340 or 343 and ends with a TAA at position 796. The corresponding polypeptide chain with the amino acid sequence indicated in Figure 3 has a calculated mol. wt. of 16 006. The amino acid composition corresponds very closely to the one determined for dUTPase by Shlomai and Kornberg (1978) (Table III). The other open reading frame starts with an ATG at position 905 and ends with TAA at position 1538. Assigned regions of interest have been listed in Table II.

Discussion

We have determined the nucleotide sequence of the *dut* gene and its flanking regions. The following points are of special interest: the sequence suggests a promoter region around position 300 with a −35 region (Rosenberg and Court, 1979) at position 286 and a Pribnow box, AATTAT, at position 310. A *Sac*I site (position 297) earlier reported to

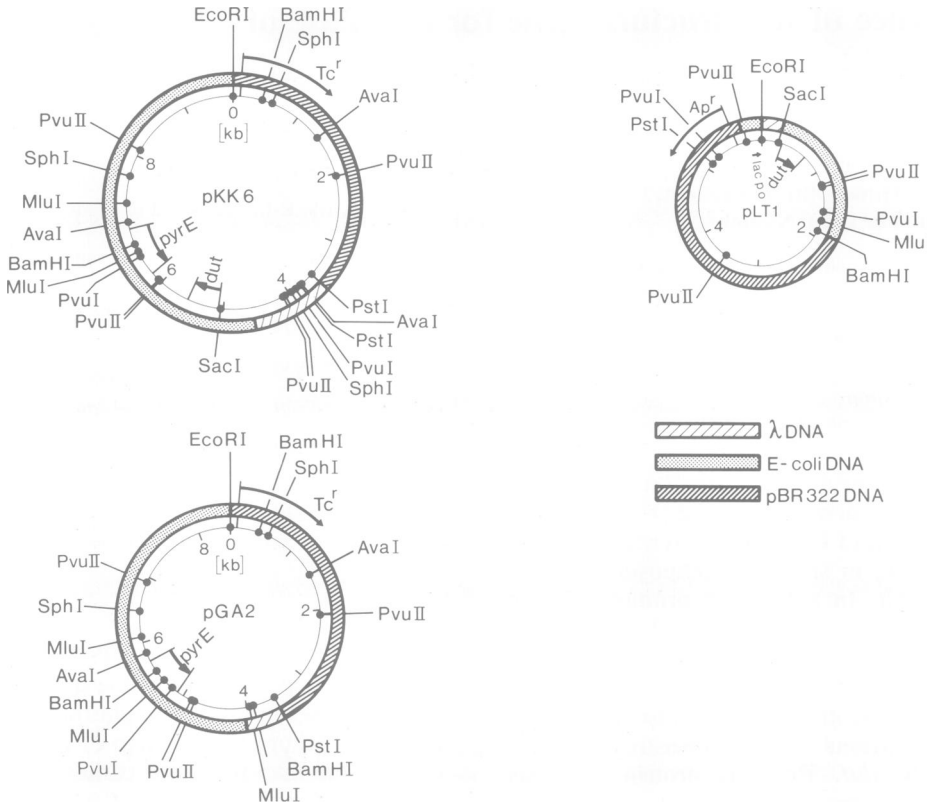

Fig. 1. Restriction endonuclease maps of plasmids utilized. Plasmid pKK6 (Lundberg *et al.*, 1979) and pGA2 (An *et al.*, 1979) have been described. These plasmids have similarly digested plasmid pBR322 (Sutcliffe, 1979) as vector. Plasmid pLT1 was constructed by cloning a 1.85-kb *Eco*RI-*Bam*HI fragment from plasmid pKK150 (Lundberg *et al.*, 1983a) in a similarly digested plasmid pBH20 (Itakura *et al.*, 1977). Bacterial strains with plasmid pKK6 overproduce dUTPase 15-fold, with pLT1 ~8-fold (unpublished data) and strains with pGA2 do not produce dUTPase (J.D. Friesen, unpublished data). The different origins of plasmid DNA are indicated. The position of various genes are shown and the directions of transcription have been indicated by arrows.

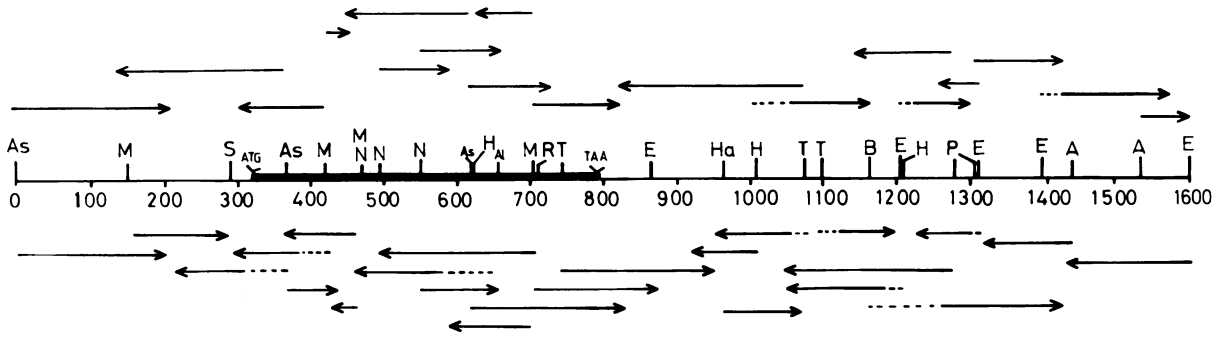

Fig. 2. Sequencing strategy for the *dut* region. Labeled ends and the extent of useful sequence information obtained from each fragment are indicated by the dots and arrows (dashed part of arrow indicates sequence not read from labeled fragment). Only sites used for labeling in this investigation are shown. For a complete restriction endonuclease map see the DNA sequence listed in Figure 3. The abbreviations of restriction endonuclease sites are: A, *Aha*III; Al, *Alu*I; As, *Asu*I; B, -*Bss*HII; H, *Hae*III; Ha, *Hae*II; E, *Eco*RII; M, *Msp*I; N, *Nci*I; P, *Pvu*II; R, *Rsa*I; S, *Sac*I; and T, *Taq*I. The *dut* gene is indicated as a black box.

be important for expression of dUTPase activity (Lundberg *et al.*, 1983a) is situated between the −10 and −35 regions. This would explain our previous observation that removal of the DNA segment upstream of the *Sac*I site strongly diminishes the expression of the *dut* gene. The corresponding transcription start would be around position 323.

The sequence shows two possible translation starts for dUTPase at position 340 and 343 (ATGATG). Calculations based on the data of Stormo *et al.* (1982), make the latter start signal the most likely one. This start signal is located 9 bp from a tentative sequence for ribosome binding (Shine-Dalgarno sequence) at position 330 (GTGA). The direction of transcription is towards the *pyrE* gene. The dUTPase gene ends at position 796 with a second stop codon at position 815. The predicted amino acid sequence and composition correlate very well with the previously published composition of the dUTPase subunit (Table III).

We have compared codon frequencies in the *dut* gene with data from other *E. coli* genes (Grosjean and Fiers, 1982). Table IV shows that very unusual codons are totally absent. Codon usage is intermediate between those of strongly and those of moderately to weakly expressed genes.

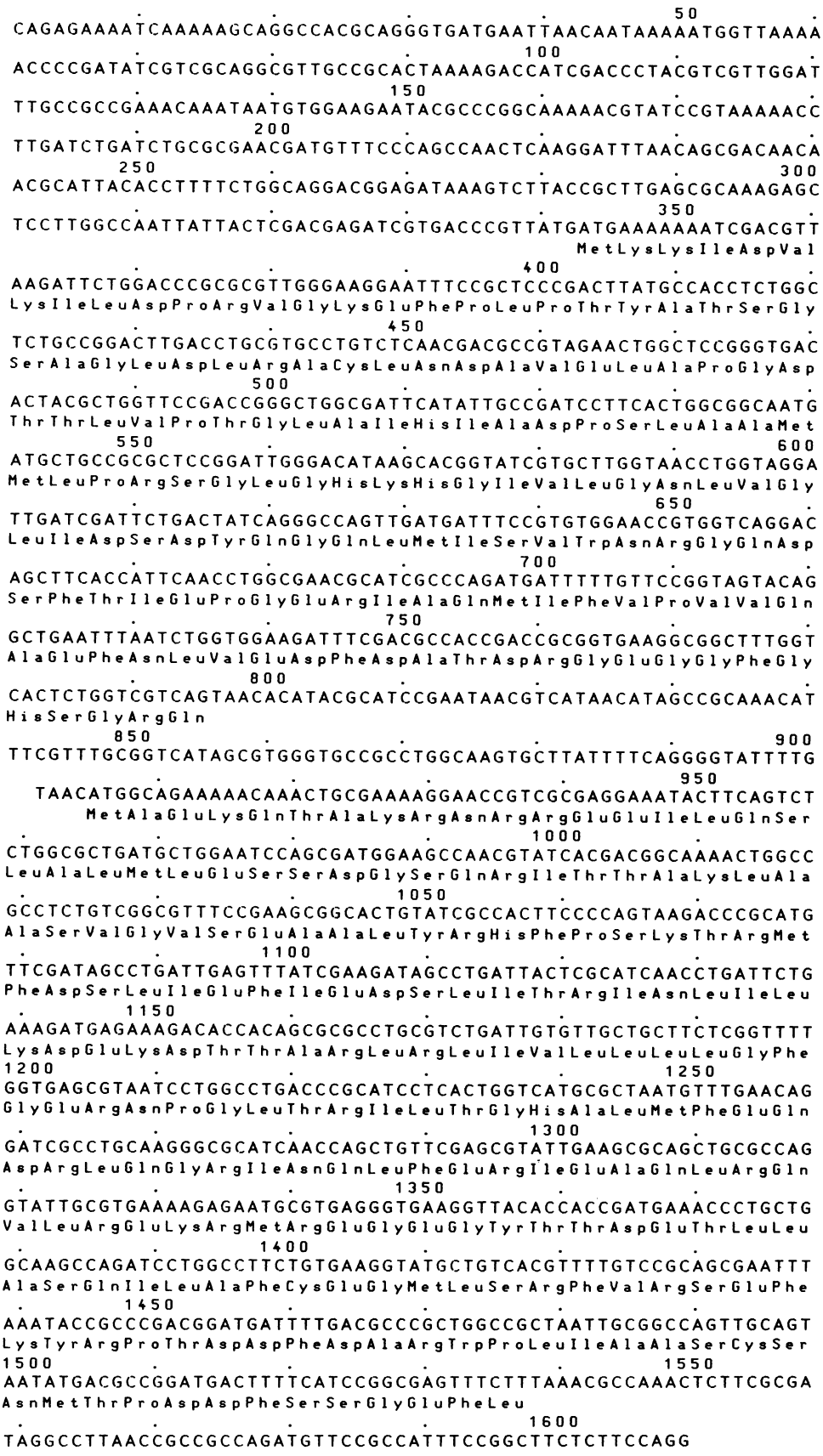

Fig. 3. DNA sequence of the *dut* region. The sequence is written in 5′ → 3′ direction of the non-coding strand. The deduced amino acid sequence for the dUTPase monomer and the ORF gene product are shown below. Special regions of interest are listed in Table II. The *pyr*E structural gene, which has an opposite direction of transcription, ends at position 1539.

The predicted amino acid sequence of dUTPase has several interesting features. It contains the sequence His-X-His which has been shown to participate in metal-binding in certain metalloenzymes (Pocker and Sarkanen, 1978; Richardson *et al.*, 1975). Since dUTPase is reported to be a zinc metalloenzyme (Shlomai and Kornberg, 1978), we suggest that this portion of the amino acid sequence could be involved in the metal binding and perhaps constitutes a part of the active site region of the enzyme. There are only one Trp and two Tyr in the dUTPase amino acid sequence which explains the low molar extinction coefficient at 280 nm for purified dUTPase (unpublished data). The polypeptide chain contains a single cysteine residue, thus excluding the occurrence of intrasubunit disulfide bridges.

The existence of a long open reading frame (ORF) starting at position 905 and ending at position 1538 is of great interest since we have previously observed the synthesis of a plasmid-coded polypeptide of mol. wt. 23 500 (Lundberg *et al.*, 1983b) which can be co-transcribed with dUTPase and which must be coded for by DNA between the *Sac*I and *Bam*HI sites. The presumed translation start region of the ORF seems reasonably similar to translation start regions of known genes. Bases 888–894 (AGGGGTA) provide a likely Shine and Dalgarno sequence.

Table II. Assigned regions of interest in the nucleotide sequence determined

	Assigned feature	Position	Sequence
The *dut* gene:			
	Promoter: −35 region	286–291	TTGAGC
	Pribnow box	310–316	AATTATT
	Shine-Dalgarno	330–333	GTGA
	Structural gene	343–795	See Figure 3
Predicted unknown gene:	Shine-Dalgarno	888–894	AGGGGTA
	Open reading frame (ORF)	905–1537	See Figure 3
Stem-loop structures:			
	Between *dut* and ORF	831–851	See Figure 3
	Between *dut* and ORF	866–893	See Figure 3

Table III. Amino acid composition of dUTPase subunit of *E. coli*

Amino acid	Number of residues		Amino acid	Number of residues	
	Nucleotide sequence	Amino acid analysis[a]		Nucleotide sequence	Amino acid analysis[a]
Ala	12	12.6	Lys	5	4.2
Arg	7	6.7	Met	4	2.7
Asp	12	16.2	Phe	6	6.5
Asn	4		Pro	9	8.8
Cys	1	0.7	Ser	8	9.7
Glu	6	14.4	Thr	7	7.1
Gln	7		Trp	1	N.D.[b]
Gly	18	18.2	Tyr	2	2.3
His	4	4.3	Val	11	10.2
Ile	10	8.9	<Total>	150	
Leu	16	15.7	Calculated mass (daltons)	16 006	

[a]Amino acid analysis of the purified dUTPase subunit (Shlomai and Kornberg, 1978).
[b]Not determined.

Table IV. Codon usage in the *dut* gene of *E. coli* K-12

Amino acid	Codon	Number of codons	Amino acid	Codon	Number of codons	Amino acid	Codon	Number of codons	Amino acid	Codon	Number of codons
Phe	UUU	4	Ser	UCU	4	Tyr	UAU	2	Cys	UGU	1
	UUC	2		UCC	2		UAC	0		UGC	0
				UCA	1						
Leu	UUA	0		UCG	0	Stop	UAA	1	Trp	UGG	1
	UUG	3									
				AGU	0						
				AGC	1						
	CUU	2	Pro	CCU	2	His	CAU	2	Arg	CGU	3
	CUC	2		CCC	0		CAC	2		CGC	4
	CUA	0		CCA	0					CGA	0
	CUG	9		CCG	7	Gln	CAA	1		CGG	0
							CAG	6			
Ile	AUU	6	Thr	ACU	2	Asn	AAU	1		AGA	0
	AUC	4		ACC	4		AAC	3		AGG	0
	AUA	0		ACA	0						
				ACG	1	Lys	AAA	2			
Met	AUG	4					AAG	3			
Val	GUU	4	Ala	GCU	2	Asp	GAU	3	Gly	GGU	7
	GUC	0		GCC	7		GAC	9		GGC	5
	GUA	4		GCA	1					GGA	4
	GUG	3		GCG	2	Glu	GAA	6		GGG	2
							GAG	0			

'Testcode' analysis (Fickett, 1982) of the ORF sequence classified the DNA segment as 'coding' (P = 0.98 and Testcode indicator, I = 1.06). The *dut* sequence when subjected to the same analysis was classified as coding (P = 1.00, I = 1.28). The predicted codon usage resembles a weakly expressed gene. We conclude from these sequence data that the ORF is likely to correspond to a structural gene. We have therefore looked for possible transcription signals in the sequence between *dut* and ORF. We have not recognized any obvious stop or start signal, but two rotational dyad symmetries are notable. One would correspond to an RNA stem-loop structure with stems from base pair 831 to 838 and 844 to 851 [ΔG = −13.1 kcal as calculated by the Tinoco rules (Tinoco *et al.*, 1973)]. Another tentative stem-loop with imperfect pairing would have stems running from base pair 866 to 872 and from 888 to 893 (ΔG = −10.4 kcal). If the presumed stem of the latter is extended to include two weaker base pairs G·T and T·A it is immediately followed by a run of four Ts. Both a GC-rich stem and a number of Ts following it is a common trait of many bacterial terminators (Rosenberg and Court, 1979), although both these characteristics are usually more pronounced. Whether termination actually occurs here to any appreciable degree can only be determined experimentally. Another feature of this region is that the second half-stem coincides almost precisely with the postulated Shine and Dalgarno region of the ORF. Perhaps the putative secondary structure plays a role in translational regulation of the expression of the ORF.

The nucleotide sequence presented here is now being extended to include the entire *pyrE* gene and its regulatory region (Poulsen *et al.*, in preparation). Data from this work rule out other locations of the gene of the 23.5-kd protein. We conclude that the ORF described (mol. wt. 24 027, predicted from nucleotide sequence), corresponds to this polypeptide of unknown function. The deduced amino acid sequence of the ORF is shown in Figure 3.

Interestingly, *pyrE* is transcribed in the opposite direction with its translation stop overlapping that of the ORF. Their transcripts are thus partially overlapping.

One other possible open reading frame co-directional with *dut* is seen in the sequence. It ends with a stop codon at position 360, thus overlapping the start of the *dut* gene. The predicted codon usage is unusual, but 'Testcode' analysis judged the polypeptide as 'coding'. If this polypeptide is a part of a protein, the translation initiation must be positioned in a region of *E. coli* DNA not covered by our *dut* plasmids. This can only be determined by sequencing *E. coli* DNA upstream from the region that we have sequenced in this investigation.

Materials and methods

Bacterial strains and plasmids

Bacterial strain NF929 (Fiil *et al.*, 1977) was transformed according to Cohen *et al.* (1972) with plasmids pLT1 (Figure 1). All strains are listed in Table I.

Growth media

The growth medium was M9 minimal medium supplemented with the required metabolites and glycerol (0.4%) as the carbon source (Miller, 1972).

DNA isolation

Plasmid DNA was purified from 500 ml bacterial cultures as described by Jørgensen *et al.* (1977). Isolation of DNA fragments from restriction enzyme digests of the plasmids was by preparative gel electrophoresis (1% agarose or 5% polyacrylamide) followed by electroelution from gel slices (McDonnel *et al.*, 1977).

Restriction endonuclease digestion

Conditions for restriction endonuclease digestion were those suggested by the manufacturer (New England Biolabs). Analysis of digestion products have been described previously (Lundberg *et al.*, 1983a).

DNA sequence determination

DNA fragments were 5′ end-labeled by [γ-^{32}P]ATP and T4 polynucleotide kinase (kindly given to us by T. Lindahl). End-label of 3′ ends was carried out by filling in sticky ends with the aid of the Klenow fragment of DNA polymerase I (New England Biolabs) with one labeled [α-^{32}P]dNTP (and the rest of the required dNTPs unlabeled), and for blunt ends with T4 DNA polymerase and [α-^{32}P]dGTP according to Challberg and Englund (1980). Labeled fragments were subjected to nucleotide sequence determination according to Maxam and Gilbert (1977). The nucleotide sequence of each strand of DNA was determined at least once. The sequences obtained were analyzed with the aid of the Queen and Korn computer programs (Queen and Korn, 1980) kindly provided by M. Johnson (University Institute of Microbiology, Copenhagen).

Chemicals

[γ-^{32}P]ATP and [α-^{32}P]dNTPs were from Amersham, UK or from NEN.

Acknowledgements

We are indebted to Dr. J.D. Friesen for sending us the strain JF1600 and to Mr. J. Skouv for making a Basic program for Testcode analyses. The expert technical assistance of Mr. P. Carlsson, Mr. P. Poulsen and Miss E.-K. Gandvik is gratefully acknowledged. This work was supported by grants from the Swedish Natural Science Research Council and Sven and Ebba-Christina Hagbergs Stiftelse.

References

An,G., Justesen,J., Watson,R.J. and Friesen,J.D. (1979), *J. Bacteriol.*, **137**, 1100-1110.
Bachmann,B.J. and Brooks Low,K. (1980) *Microbiol. Rev.*, **44**, 1-56.
Bertani,L.E., Häggmark,A. and Reichard,P. (1961) *J. Biol. Chem.*, **236**, 67-68.
Challberg,M.D. and Englund,P.T. (1980) *Methods Enzymol.*, **65**, 39-43.
Cohen,S.N., Chang,A.C.Y. and Hsu,L. (1972) *Proc. Natl. Acad. Sci. USA*, **69**, 2110-2114.
Fickett,J.W. (1982) *Nucleic Acids Res.*, **10**, 5303-5318.
Fiil,N.P., Willumsen,B.M., Friesen,J.D. and von Meyenberg,K. (1977) *Mol. Gen. Genet.*, **150**, 87-101.
Grosjean,H. and Fiers,W. (1982) *Gene*, **18**, 199-209.
Itakura,K., Hirose,T., Crea,R., Riggs,A.D., Heyneker,H.L., Bolivar,F. and Boyer,H.W. (1977) *Science (Wash.)*, **198**, 1056-1063.
Jørgensen,P., Collins,J. and Valentin-Hansen,P. (1977) *Mol. Gen. Genet.*, **155**, 93-102.
Lundberg,L., Nyman,P.O. and Karlström,O. (1979) *Acta Chem. Scand*, **B33**, 597-598.
Lundberg,L.G., Karlström,O.H., Nyman,P.O. and Neuhard,J. (1983a) *Gene*, **22**, 115-125.
Lundberg,L.G., Karlström,O.H. and Nyman,P.O. (1983b) *Gene*, **22**, 127-131.
Maxam,A. and Gilbert,W. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 560-564.
McDonnel,M.W., Simon,M.N. and Studier,F.W. (1977) *J. Mol. Biol.*, **110**, 119-146.
Miller,J.H. (1972) *Experiments in Molecular Genetics*, published by Cold Spring Harbor Laboratory Press, NY.
Pocker,Y. and Sarkanen,S. (1978) *Adv. Enzymol.*, **47**, 149-274.
Queen,C.L. and Korn,L.J. (1980) *Methods Enzymol.*, **65**, 595-609.
Richardson,J.S., Thomas,K.A., Rubin,B.H. and Richardson,D.C. (1975) *Proc. Natl. Acad. Sci. USA*, **72**, 1349-1353.
Rosenberg,M. and Court,D. (1979) *Annu. Rev. Genet.*, **13**, 319-353.
Shlomai,J. and Kornberg,A. (1978) *J. Biol. Chem.*, **253**, 3305-3312.
Stormo,G.D., Schneider,T.O., Gold,L. and Ehrenfeucht,A. (1982) *Nucleic Acids Res.*, **10**, 2997-3011.
Sutcliffe,J.G. (1979) *Cold Spring Harbor Symp. Quant. Biol.*, **43**, 77-90.
Taylor,A.F., Siliciano,P.G. and Weiss,B. (1980) *Gene*, **9**, 321-336.
Tinoco,I.,Jr., Borer,P.N., Dengler,B., Levine,M.D., Uhlenbeck,O.C., Crothers,D.M. and Gralla,J. (1973) *Nature, New Biol.*, **246**, 40-41.
Tye,B.K., Nyman,P.O., Lehman,I.R., Hochhauser,S. and Weiss,B. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 154-157.