Genetic Regulation: The Lac Control Region

The nucleotide sequence of the lac control region containing the promoter and operator is presented.

Robert C. Dickson, John Abelson,
Wayne M. Barnes, William S. Reznikoff

The molecular basis of gene regulation is best understood in the case of the lactose (lac) operon in *Escherichia coli*. When inducers such as lactose or other β-galactosides are added to a culture of *E. coli*, there is a 1000-fold increase in the rate of synthesis of β-galactosidase, β-galactoside permease, and thiogalactoside transacetylase. The genes for these three proteins are linked together on the *E. coli* chromosome as shown in Fig. 1. The three genetic elements involved in control of the lac operon are also shown. These are the repressor gene, *i*, the promoter, *p*, and the operator, *o*. It was the phenotypic properties of lac *i* and lac *o* mutations that led Jacob and Monod to propose their classic model for the control of gene expression (*1*). An updated version of this model as it applies to the lac operon is reviewed below.

1) Genes function by serving as templates for the synthesis (transcription) of messenger RNA (mRNA). The complex protein synthesizing machinery translates mRNA into polypeptide chains.

2) The operon, genes *z*, *y*, and *a*, is a single unit of transcription. Transcription is initiated at *p*. Mutations in *p* result in a coordinate change in the level of synthesis of all products of the operon (*2*).

3) Transcription of the operon is both negatively and positively controlled. Negative control is mediated by the lac repressor which binds specifically and tightly to *o* thereby preventing transcription (*3*). Mutations in *o* or in the *i* gene generally result in constitutive synthesis of the products of the operon (*1*). Positive control of the lac operon is exerted via the phenomenon termed catabolite repression. Expression of the lac operon (and other catabolite repressible operons) is repressed when glucose, a more efficient source of carbon than lactose, is present in the medium. By an as yet undetermined mechanism, the presence of glucose results in a decreased concentration of intracellular adenosine 3′,5′-monophosphate (cyclic AMP) (*4*). Cyclic AMP is required for efficient expression of the lac operon since it activates the catabolite gene activator protein (CAP) which in turn activates transcription of lac by RNA polymerase (*5*).

Control of gene expression in the lac operon is thus mediated by the specific interaction of three proteins, RNA polymerase, CAP, and the lac repressor, with distinct DNA sequences in the lac *p-o* region as shown in Fig. 1. These sequences have been defined genetically; the function of each sequence can be inferred from the phenotype of mutations which alter the sequences. Locating these mutations in the lac *p-o* nucleotide sequence could allow us to assign functions to specific portions of the sequence. Moreover, knowledge of the sequence would be a first step toward a detailed physical-chemical description of the interactions between these three proteins and their DNA recognition sites. In this article we report the determination of the nucleotide sequence for the lac *p-o* region.

The sequence was determined by analysis of RNA transcripts of the lac *p-o* region. The method used to isolate this RNA (*6*) is summarized in Fig. 2. The method requires (i) uniform transcription of the lac *p-o* region in vitro either in the "natural" (LAC) direction or in the reverse (CAL) direction; (ii) specialized λ transducing phages carrying genetic deletions defining both sides of the *p-o* region; and (iii) specialized λ transducing phages carrying the *p-o* DNA in opposite orientations vis-à-vis the phage genes. The construction of these phages has been described (*7*). An advantage of this procedure is that it does not rely on the physiological activity of the genetic control signals under study. It can, therefore, be applied without modification to the study of point mutations introduced into the DNA template.

The standard two-dimensional paper electrophoresis (oligonucleotide map) techniques developed by Sanger and coworkers (*8*) were used for separating oligonucleotides. RNA was labeled in vitro with one nucleoside [α-^{32}P]triphosphate. Figure 3A shows a typical ribonuclease T1 oligonucleotide map of guanosine [α-^{32}P]triphosphate–labeled CAL RNA whose ends are defined by deletions X8630 and X8555. This result shows that the product is a unique RNA molecule with a complexity of about 140 nucleotides. This RNA is complementary to *p* and *o*, as can be shown by the effects of introducing various deletions in the *p-o* region of the DNA used for RNA-DNA hybridization. For example, Fig. 3B shows a ribonuclease T1 oligonucleotide map of CAL RNA whose ends are defined by deletions X8554 and W227. Clearly several oligonucleotides have been removed by these deletions. Analytical hybridization and further separation data have proved the specificity of the *p-o* RNA (*6*). Figure 3C shows a comparable ribonuclease T1 oligonucleotide map of LAC RNA. This pattern is unique and has about the same complexity as CAL RNA.

Determination of the sequence. The nucleotide sequences (Fig. 4) of CAL and LAC RNA were determined in the following way.

1) Oligonucleotides produced by separate digestions of LAC and CAL RNA with either pancreatic ribonuclease A or ribonuclease T1 were sequenced by nearest neighbor sequencing techniques (*8*). All but the five largest oligonucleotides, CAL T25, CAL T26, LAC T17, LAC T19, and LAC T20, could be sequenced from the nearest neighbor data alone (Tables 1 and 2). Diges-

Dr. Dickson is an assistant research chemist and Dr. Abelson is an associate professor in the Department of Chemistry at the University of California, La Jolla 92037. Dr. Barnes was a graduate student and Dr. Reznikoff is an associate professor in the Department of Bicchemistry, College of Agriculture and Life Sciences, University of Wisconsin, Madison 53706.

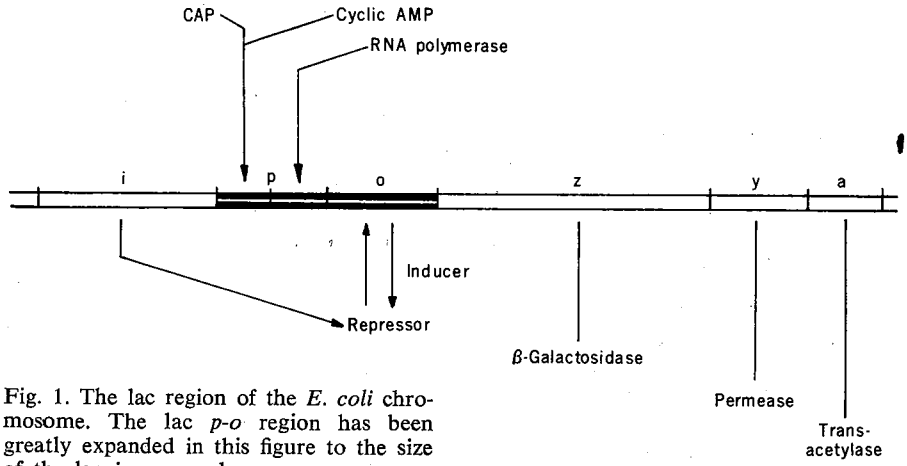

Fig. 1. The lac region of the *E. coli* chromosome. The lac *p-o* region has been greatly expanded in this figure to the size of the lac *i*, *z*, *y*, and *a* genes.

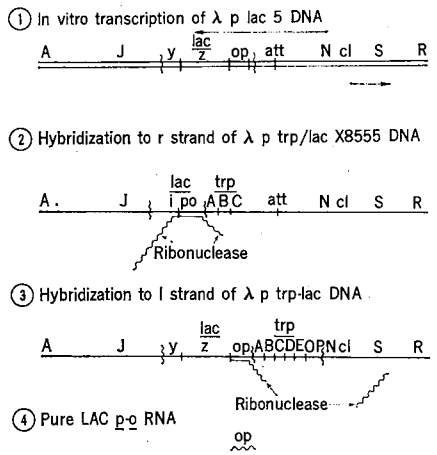

Fig. 2. Isolation of RNA complementary to *p* and *o*. RNA is transcribed in vitro from lambda (λ) *plac* 5 DNA with *E. coli* RNA polymerase holoenzyme. Our conditions in vitro allow transcription of sequences that are not transcribed in vivo such as the lac promoter. RNA complementary only to the *p-o* region is isolated by first hybridizing the complete in vitro transcript to the r strand of λ*ptrp/lac* X8555 DNA. The lac genes *z*, *y*, and *a* are deleted from λ*ptrp/lac* X8555. All unhybridized RNA including that complementary to genes *z*, *y*, and *a* and to the l strand of λ is digested with ribonuclease A leaving a piece of hybridized RNA that has one end terminating near the start of the *z* gene. This RNA is eluted and rehybridized to the l strand of λ*ptrp-lac* X8630, a transducing phage in which most of the lac *i* gene has been deleted. Unhybridized RNA, particularly that complementary to the *i* gene and to the r strand of λ, is digested with ribonuclease A, and the hybridized RNA is eluted. We call this LAC RNA since it has been transcribed in the same direction as lac mRNA. By using φ80*plac* I DNA as the in vitro template and by using opposite DNA strands for the RNA-DNA hybridizations we can isolate RNA transcribed from the opposite strand. We call this CAL RNA. More than 90 percent of the final RNA hybridizes specifically to the *p-o* region (*6*).

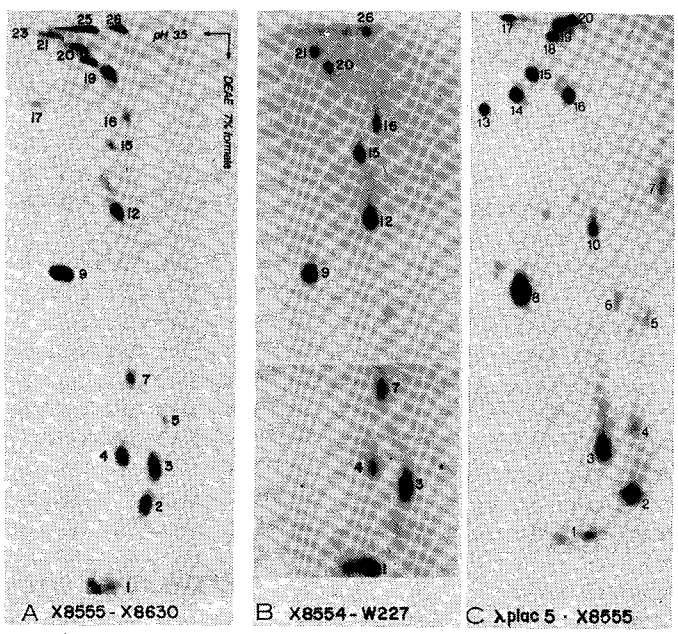

Fig. 3. Two-dimensional ribonuclease T1 oligonucleotide maps of RNA complementary to the lac *p-o* region. These autoradiographs show the oligonucleotide maps of CAL RNA labeled with guanosine [α-^{32}P]triphosphate, panels A and B, and LAC RNA, panel C, prepared as described in Fig. 2. The complexity of the RNA is determined by the deletions (Fig. 5) indicated below each panel. The sequence of each numbered oligonucleotide is presented in Table 1. A comparison of A and B shows that deletions X8554-W227 remove several oligonucleotides present in the X8630-X8555 set and demonstrate how oligonucleotides can be positioned in a segment (Fig. 5) of the lac *p-o* region.

tion of T17 with ribonuclease U2—this enzyme yields products ending in 3′ G (guanosine) or A (adenosine)—gives 2 moles of CUUUA (C, cytidine; U, uridine). Of the ten possible sequences compatible with the nearest neighbor data, only the one in Table 1 is compatible with this result. Digestion of LAC T19 with ribonuclease U2 gives 2 moles of CUCA. Of 14 possible sequences compatible with the nearest neighbor data, two sequences could yield this result. Only the sequence given in Table 1 is compatible with the sequences of oligonucleotides in the complementary strand. The sequence of CAL T25 is determined by sequences in the complementary strand. Ambiguities remain in the sequence of CAL T26 (Table 5) and LAC T20 (Table 1). The sequence of CAL T19B was not independently determined because accurate nearest neighbor data could not be obtained (Table 1).

2) The lac deletions carried by λ phages allowed us to divide the sequence into six segments (I to VI, Fig. 5). By examining RNA whose ends are defined by various pairwise combinations of these deletions we could place many of the ribonuclease A and T1 oligonucleotides in a particular segment (last column of Tables 1 and 2). This reduced the complexity of the sequencing task. Essentially we were faced with the problem of sequencing three overlapping blocks (segments I, II, and III; segments III, IV, and part of V; and segments V, VI, and part of IV) of about 50 nucleotides each, rather than the more difficult job of sequencing one molecule of 140 nucleotides.

3) The nearest neighbor data (Tables 1 and 2) give a large number of possible sequences for each block. Overlaps between the ribonuclease A and T1 digestion products for a particular strand restrict the number of possible sequences but do not uniquely determine the sequence for any block. Assuming Watson-Crick base pairing, we have used oligonucleotides from the complementary strand to obtain information on the ordering of the T1 products. This is a powerful constraint and in general severely limits the number of possible sequences. Ambiguities in the sequence were further resolved by the isolation and sequence analysis of partial ribonuclease A digestion products of LAC RNA. Several useful overlaps in block 1 and block 3 were obtained in this manner (Tables 3 and 5, respectively).

Table 1. Ribonuclease T1 products from CAL RNA and LAC RNA. Ribonuclease T1 oligonucleotides obtained from CAL and LAC RNA. Oligonucleotide numbers refer to those shown in Fig. 3, A and C. Nearest neighbors are given in brackets. In some cases, for example, the ribonuclease A product AAU[U] from T25, the nearest neighbor was obtained from an alkali digest of the product. Products are normalized to 1 mole unless preceded by another integer. Roman numerals indicate in which segment (Fig. 5) of the sequence an oligonucleotide belongs. The letters NL mean that a product did not become labeled by the indicated radioactive triphosphate; R, treatment with ribonuclease A; Alk, treatment with 0.5*M* NaOH.

Oligonucleotide	Moles (No.)	Treatment	Labeled products resulting from α^{32}-P labeled ATP	CTP	GTP	UTP	Deduced sequence and position (segment No.)
			Ribonuclease T1 products from CAL RNA				
T1	4	R	G	NL	G	G	2 G[G]*IV; G[U]IV; G[A]IV
T2	2	Alk	NL	G	C	G	CG[U]*I + II; CG[C]I + II
T3	1	R	NL	C	C, G	NL	CCG[G]*IV
T4	3	Alk	NL	G	A	G	AG[C]*IV; AG[C]III; AG[U]III
T5	1	R	NL	C	C	NL	CCCG[C]†I + II
T7	1	R	AAG	AAG	AAG	NL	AAG[C]*IV
T9	5	Alk	G	G	U	G	UG[A]*V + VI; UG[A]III; UG[C]IV; UG[U]IV; UG[U]V + VI
T10	1	R	NL	NL	U	C, G	CUG[U]*VI
T12	1	R	NL		U, G		CCUG[G]*IV
		Alk	NL	C	U, G	C	
T15	1	R	AAAG, U		AAAG	NL	UAAAG[C]IV
		Alk		G	A	NL	
T16	1	R	AAAG, AU, C	NL	AAAG	AAAG, AU	CAUAAAG[U]IV
		Alk		NL	A	A, G	
T17	1	R	NL	G	U	U	UUG[C]*I + II
T19A	1	R	AAU, G, U	C	AAU	AAU, C	CCUAAUG[A]III
T19B	1	R	C	AC, G, U	U	AC, C	CUCACUG[C]‡I + II
T20	1	R	AAAU	NL	U	AAAU[U], G	AAAUUG[U]V
T21	1	R	U	AU, G, C	C	AU, U	UUAUCCG[C]V
T23	1	R	NL	C, U	U	G, C, 2U	UUUCCUG[U]§VI
T25	1	R	AAU, AAC, AC, C, 2U	AAC, AC, U, G	U	AAU[U], AAC, AU[U], C	. . . UG[C]¶I + II
T26	1	R	AAU, AAC[A], AU, 3AC, G, 2C	AAC, 4AC, C, 2U	AC	AAU[U], AU, C	. . . ACG[A]**IV − V
			Ribonuclease T1 products from LAC RNA				
T1	7–8	R	G	G	G	NL	3 G[C]*IV; G[C]I + II; G[G]I + II; G[A]IV − V; 2G[A]V
T2	3	R	NL	G	G, C	NL	2 CG[G]*V; I + II; CG[C]I + II
T3	3	R	NL	AG	AG	AG	AG[C]*I + II; AG[C]V; AG[U]III
		Alk			A		
T4B	1	R	C	NL	AG	AG	CAG[U]*I, II
T4A	<1	R	AAAC	AAAC	NL	NL	AAAC††VI
T5	1	R	AAC, C	AAC,G	AAC	NL	CAACG[C]I + II
T6	1	R	AAAG	AAAG	AAAG	NL	AAAG[C]I + II
T7	1	R	C	AC, C	AG	NL	CACCCCAG[G]‡‡IV
		Alk			A, G		
T8	5	R	G	NL	U,G	G	2 UG[A]*I + II; UG[A]V; UG[G]; UG[U]IV − V
T10	1	R	NL	U	C	C, G	CUCG[U]IV
T13	2	R	NL	NL	U	G, U	UUG[U]*IV − V
T14	2	R	U	AG	AG, AU	AU, G, U	UAUG[U]IV − V; UUAG[C]III
T15	1	R	AAU	NL	U	AAU[U], G	AAUUG[U]IV − V
T16	1	R	NL	C, U	G, C	C, U	CUUCCG[G]IV or CCUUCG[G]§§
T17	1	R	AC, 2U	2AC, G	AU	AU, AC, C, 4U	CUUUACACUUUAUG[C]¶¶IV
T18	1	R	AAU, C, U	NL	AAU	AAU[U], G	CAAUUAAUG[U]I + II
T19	1	R	U, 2C	AC, 2U	AG	AU[U], AC, C	CUCACUCAUUAG[G]¶¶III
		Alk			A, G		
T20	1	R	AAU, AAC[A], 2AC, AU, C	AAC, 2AC, U	AG	AAU[U], AU, U	. . . AG[G]¶V
		Alk			A, G		

* Composition determined in part from mobility on a two-dimensional oligonucleotide map (Fig. 3). † The mobility of this spot indicates three C's. ‡ This sequence is dictated by complement, oligonucleotide LAC T4B, and P12 as shown in Fig. 4. The nearest neighbor data for CAL T19B are weak because the oligonucleotide was rarely present in molar yield and did not reproducibly separate from CAL T19A. § Of three possible sequences only this one is complementary to LAC oligonucleotide P14. ¶ This oligonucleotide has not been sequenced. **This sequence is discussed in the legend to Table 5. †† The presence of this product in a ribonuclease T1 digest indicates that the ribonuclease A used in the purification procedure often cleaved at this point. ‡‡ Of three possible sequences only this one is complementary to CAL oligonucleotide P18. §§ This sequence is eliminated because there is no GAAGG sequence in the CAL strand. ¶¶ The derivation of this sequence is discussed in the section on sequencing, step 2, of the text.

4) To prove that the sequence for each block is unique we have employed a computer program (*9*), which generates all possible sequences consistent with a set of T1 products and their nearest neighbors. Each member of the set is then compared with the ribonuclease A digestion products from the same strand and both ribonuclease A and T1 digestion products from the opposite strand. The data for blocks 1, 2, and 3 and the resulting sequences (Tables 3, 4, and 5), indicate that the sequence for block 1 is unique. There is an unresolved ambiguity in the se-

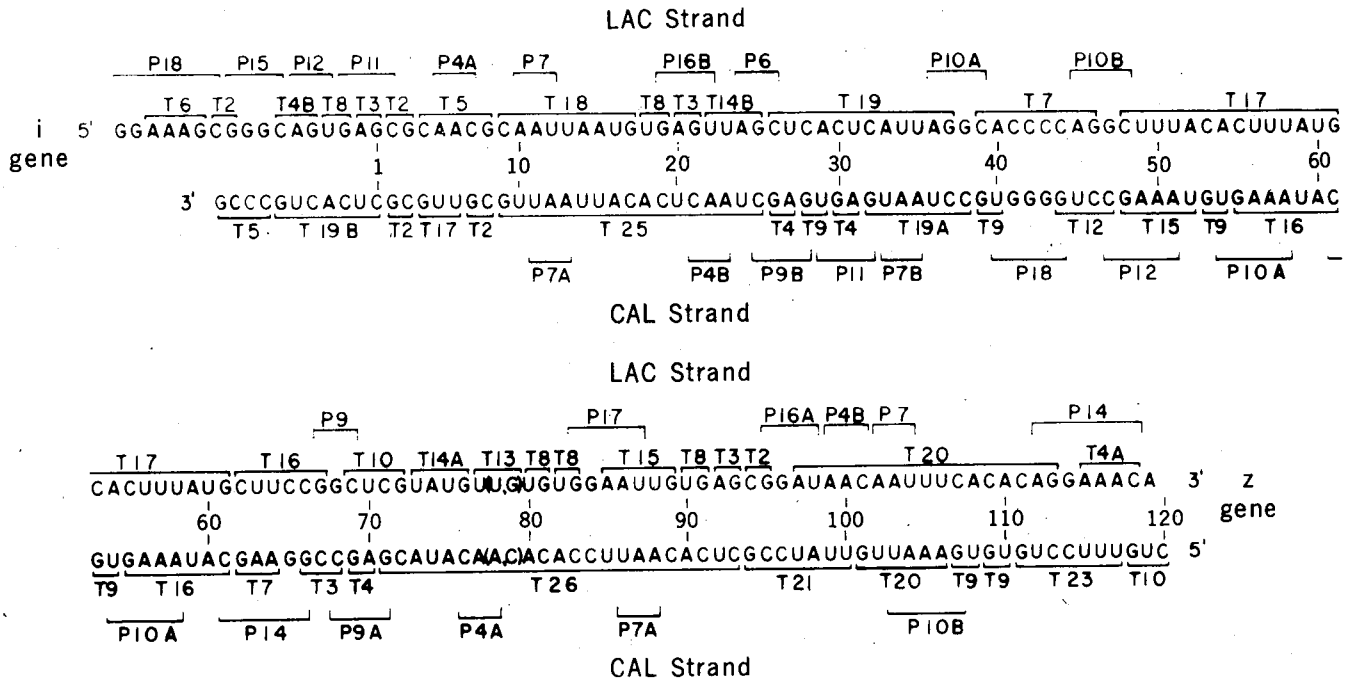

Fig. 4. LAC and CAL *p-o* RNA sequences. The LAC and CAL *p-o* RNA sequences are presented with the significant T1 oligonucleotides (labeled T) and ribonuclease A oligonucleotides (labeled P). The numbers between the two strands indicate individual nucleotide pairs starting with the first nucleotide beyond the lac *i* gene stop codon.

Table 2. Some ribonuclease A digestion products from CAL RNA and LAC RNA. Oligonucleotides were separated by two-dimensional electrophoretic mapping (see Table 1).

Oligonucleotide	Moles (No.)	Treatment	Labeled products resulting from α-^{32}P labeled				Deduced sequence and position (segment No.)
			ATP	CTP	GTP	UTP	
			Ribonuclease A products from CAL RNA				
P7		Alk	A	NL	2A, U		AAU[G]*II; AAU[U]
		T1			AAU		
P8		Alk	U	NL	U	U, 3G	GU[A]*; GU[G]; GU[U]
P9	2	T1	G	AG, C	AG	C	GAGC[C]IV; GAGC[U]III
P10	2	T1	AAAG, G, AAAU	NL	AAAG, U	AAAU, AAAG	AAAGU[G]IV; GAAAU[U]V
		Alk			A, U	A, G, U	
P11	1	T1	G	NL	AG, U	AG	GAGU[G]*III
P12	1	T1	AAAG	AAAG, C	AAAG	NL	AAAGC[C]IV
P14	1	T1	AAG, G, C	AAG	AAG, C	NL	GGAAGC[A]IV
P18	1	T1	NL	NL	G, U	G	GGGGU[G]†IV
			Ribonuclease A products from LAC RNA				
P6		T1	NL		AG	C	AGC[U]*III
		Alk		G		C	
P7		Alk	A	NL	U	A, U	AAU[G]*II; AAU[U]V
P8		Alk	U	NL	U	U, G	GU[A]*; GU[G]; GU[U]
P9	1	T1	NL	G	G	C	GGC[U]*IV
P10, P11	4	T1	G, C	AG, G	AG[G], AG, C	C	AGGC[A]*III + IV; AGGC[U]IV; GAGC[G]V; GAGC[G]; I + II
P12	1	T1	NL	NL	AG, U	AG	AGU[G]*I + II
P14	1	T1	AAAC, G	AAAC	AG	NL	AGGAAAC VI
P15	1	T1	C	G	G	NL	GGGC[A]*I + II
P16	2	T1	AU[A], G	NL	AG, G	AG, AU, U	GGAU[A]V; GAGU[U]III
P17	1	T1	AAU, G	NL	G	AAU[U]	GGAAU[U]IV – V
P18	< 1	T1	AAAG, G	AAAG	AAAG, G, C	NL	GGAAAGC[G]I

* Composition was also determined by mobility on a two-dimensional oligonucleotide map. † The number of G's was determined by the complement strand, oligonucleotide LAC T7, and also by mobility on a two-dimensional oligonucleotide map.

quence for block 2 involving LAC T13-UUG(U) and LAC T8-UG(U) at nucleotides 77 to 81 (Table 4). We have not ordered these oligonucleotides; consequently, we do not know the sequence for nucleotides 78 and 79 (Table 4 and Figs. 4 and 5). This ambiguity could be resolved by independently sequencing CAL T26. The two possible sequences for block 3 differ in nucleotides 101 to 108. One sequence (in CAL RNA) is UGAAAUUG and the other is AAAUUGUG. The former agrees with the operator sequence of Maizels (*10*) and Gilbert and Maxam (*11*), and we assume it is the correct sequence. To verify this sequence we will have to determine the sequence of LAC T20. The RNA sequence for the entire *p-o* region is shown in Fig. 4. The position of significant T1 and ribonuclease A products is indicated.

The complete DNA sequence of the lac *p-o* region defined by the end points of the deletions carried by *λptrp/lac* X8555 and *λplac* 5 is shown in Fig. 5. The sequence overlaps on the left with the *i* gene and on the right it comes close to the *z* gene. Between the UGA stop codon in the *i* gene and the AUG start codon in the *z* gene, there are 122 base pairs (66 A · T and 56 G · C).

Overlap with the i *gene.* The amino acid sequence of the *i* gene product, the lac repressor, is known (*12*). A possible codon sequence for the four amino acids at the carboxyl-terminus correlates with the sequence of the DNA at the 5′ end of the lac strand (Fig. 5). There is a UGA stop codon that immediately follows the carboxyl-terminus of the glutamine (Gln) codon. In addition there is another stop codon in phase with the first one but four codons away from it.

The sequence in this region extends further into the *i* gene on the LAC strand than on its complement (Fig. 4). This may be due to the proportionately high number of purines on the LAC strand in this region. These sequences would be resistant to the ribonuclease A used in trimming the hybrids. This observation emphasizes that the end points of the deletions in the *λptrp-lac* phages have not been determined exactly.

Experiments by Lebowitz *et al.*, Pieczenik *et al.*, Dahlberg and Blattner, and Ikemura and Dahlberg (*13*) suggest that the sequence TTTTTTA is an mRNA transcription termination signal both in vivo and in vitro. There is no such signal at the end of the *i* gene. Recent experiments have indicated that there may be no mRNA stop signal for the *i* gene in vivo (*14*), and our own experiments [those reported here and in (*6*)] indicate that there is no obligatory transcription stop signal in vitro.

The o-z *region.* Our sequence ends two to four nucleotide pairs before the initiation codon for β-galactosidase (see Fig. 5). We infer this since our sequence prior to the initiation codon agrees with the sequence of lac mRNA determined by Maizels (*10*). Her sequence extends eight codons into the *z* gene and agrees with that predicted from the amino-terminal amino acid sequence of β-galactosidase.

The operator. The operator is defined genetically by a set of operator constitutive (o^c) mutations that result in *cis*-dominant constitutive expression of the operon (*1*). Binding studies in vitro of repressor to DNA (*3*) show that these mutations decrease the affinity of repressor for the operator. In our sequence, the operator is defined by the end points of deletions X8554 and S20 because *λptrp/lac* X8554 DNA binds repressor with a high affinity, whereas *λptrp/lac* F36a-S20 and *ϕ80plac* S20 DNA show no binding (*15*). The operator is distinguished by a region of partial twofold rotational symmetry (Fig. 5) as has been noted (*11*). Our results (Fig. 5) support and extend this symmetry by an additional four to six base pairs. The function of this rotational symmetry is unknown, but there are several possibilities.

1) From the standpoint of evolutionary economics, the protein partner in a symmetrical interaction of protein and DNA requires informational content only half that of an asymmetric interaction with a DNA recognition site of similar size, specificity, and binding energy.

2) The symmetrical sequences permit a protein that is capable of one-dimensional diffusion along the DNA molecule to recognize the desired sequence from either direction.

3) The symmetrical sequences permit the generation of looped cruciform structures, as suggested by Gierer (*16*), which could be recognized by proteins. It seems unlikely on energetic grounds that such a structure exists in DNA in the absence of a stabilizing protein interaction. Studies by Wang, Barkley, and Bourgeois (*17*) show that a cruciform structure does not form when repressor binds to the operator.

The promoter. The promoter is a region of DNA required for transcription initiation (*2*). In the lac operon the promoter is genetically defined by mutations that alter the maximum level of gene expression in the operon. These mutations all map between the ends of deletions X8630 and X8554 and presumably are located between X8630 and W225 (*18*). The promoter can be divided into two functional units, the CAP interaction site and the RNA polymerase interaction site (*19*) (Figs. 5 and 6).

The CAP interaction site. The CAP interaction site is defined by class I promoter mutations (*18, 19*). These mutations reduce expression of the lac operon presumably by lowering the affinity of the promoter for CAP. The deletion L1 is a class I mutation, and

Table 3. Computer analysis of sequence in block 1 (segments I, II, and III). Oligonucleotide designations refer to those given in Tables 1 and 2 and Fig. 4 except for oligonucleotides designated MPP. These were produced (*30*) by partially digesting CAL or LAC RNA with pancreatic ribonuclease carboxymethylated in lysine 41. The resulting oligonucleotides were digested with ribonuclease A or T1 and analyzed by standard sequencing procedures (data not shown). If known, the nearest neighbor for an oligonucleotide is included in this table and in Tables 4 and 5. To obtain the LAC sequence shown at the bottom of this table, the computer generated all sequences compatible with the products listed under the column headed LAC T1 oligonucleotides. Sequences not compatible with the LAC ribonuclease A oligonucleotides and the CAL ribonuclease A and T1 oligonucleotides listed in the second and third column were rejected by the computer. Only the sequence shown fits these data.

	LAC T1 oligonucleotides
T4B	CAGU
T8	2UGA
T18	CAAUUAAUGU
T3	AGU
T14B	UUAGC
T19	CUCACUCAUUAGG
MPP	GAGCGCAACGC (composed of T3, T2, and T5)
MPP	GGAAAGCGGGC (composed of P18 and P15)
	LAC ribonuclease A oligonucleotides
P16B	GAGUU
P6	AGCU
P12	AGUG
	CAL T1 and ribonuclease A oligonucleotides
T19A	CCUAAUGA
P11	GAGUG
T17	UUGC
Part of:	
T25	AAUU
T25	UGC
T25	AAC

LAC sequence

GGAAAGCGGGCAGUGAGCGCAACGCA

— 10 1 10

AUUAAUGUGAGUUAGCUCACUCAUUAG

20 30

all other class I mutations map within this deletion. The deletions F23a, F36a, and W227 also prevent CAP stimulation of lac expression (*14*). Therefore, if the CAP interaction site does not overlap with the *i* gene [as suggested by the phenotype of deletion X8630 (*20*)], it can be localized in a sequence approximately 37 base pairs long between the *i* gene stop signal and the left end of L1. Sixteen of these base pairs are in a region of twofold rotational symmetry, which we tentatively conclude is the recognition site for CAP (see Figs. 5 and 6), a protein composed of two identical subunits each having a molecular weight of 22,000 (*21*). Confirmation that this symmetrical sequence is indeed recognized by CAP requires sequence analyses of class I point mutations.

The RNA polymerase interaction site. The interaction site for RNA polymerase is defined by the deletion L1, which does not affect CAP independent lac expression, and the nucleotide that codes for the 5′ end of the lac mRNA

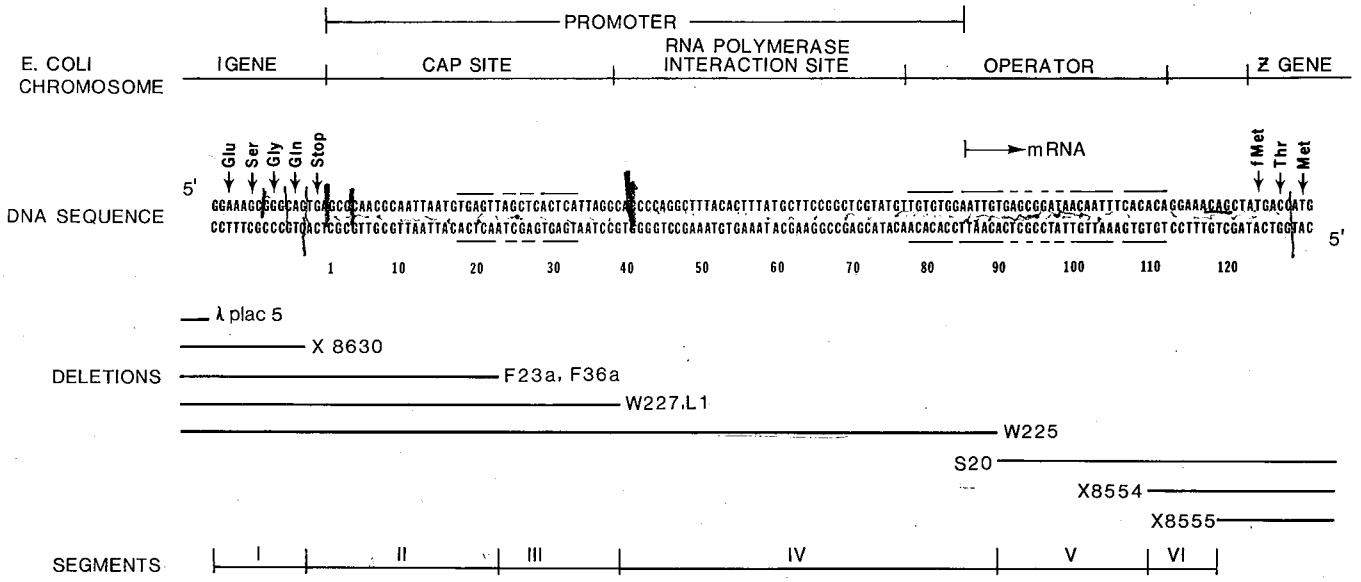

Fig. 5. The lac *p-o* DNA sequence. The LAC and CAL RNA sequences have been changed into a DNA sequence with the sequence extended 11 base pairs to include the first three codons of the *z* gene [data from Maizels (*10*)]. Indicated above the sequence are the proposed locations of the *i* gene, *p* (CAP and RNA polymerase sites), *o*, and *z* genes. The regions of symmetry in the CAP site and *o* are shown as is the mRNA start site which has been determined by Maizels and by Majors (*10, 22, 31*). Below the sequence, the approximate locations of the deletions used in the RNA hybridization procedures are shown as well as the segments defined by some of these deletions.

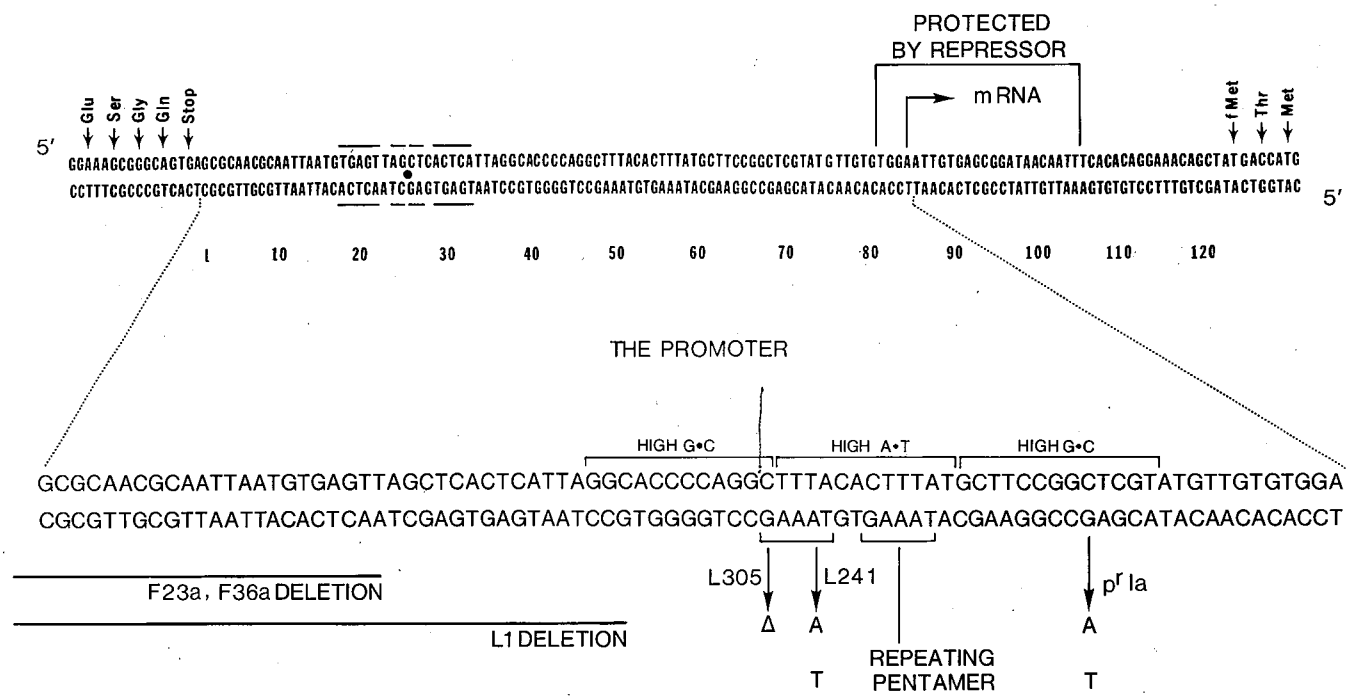

Fig. 6. The lac promoter sequence. The detailed sequence of the lac promoter shows the location of the various promoter mutations and interesting structural features such as the CTTTA repeats, the CAP symmetry elements, and the high G · C and A · T blocks. The class II mutation L305 deletes a G · G base pair. The sequence protected by the repressor was determined by Gilbert and Maxam (*11*).

as determined in vitro by Maizels (*10*) and Majors (*22*) (Figs. 5 and 6). This region is approximately 46 base pairs long. Located within this sequence are two class II promoter mutations (L241 and L305) which decrease CAP independent expression in the lac operon (*18*). The nucleotide changes produced by these two mutations are indicated in Fig. 6. This region contains no symmetry elements involving more than four base pairs, suggesting that sequence symmetry is not important in the interaction between RNA polymerase and DNA. This is consistent with the unidirectional nature of mRNA synthesis. As is shown in Fig. 6 this sequence contains the following interesting structural features:

1) A repetitious sequence, which on the LAC strand reads CTTTA (T, deoxyribothymidine) (*23*). The two class II mutations are located in the first repetition.

2) Sequence blocks with unusually high G · C and A · T content.

3) A similarity between part of this sequence and that determined for one of the phage fd RF promoters by Schäller *et al.* (*24*). The end of the piece of fd RF DNA protected from deoxyribonuclease digestion by RNA polymerase contains a sequence of six base pairs in common with the lac promoter (nucleotides 60 to 65 in Fig. 6) followed by a region in which 8 of 20 base pairs are identical (nucleotides 67, 69, 70, 74, 75, 81, 83, and 85 in Fig. 6).

It should be noted that the proposed promoter sequence overlaps (by up to eight base pairs) with the region of symmetry containing the operator.

Initiation of transcription—a model. In vitro studies on the initiation of RNA transcription suggest that this process can be divided into four or possibly five steps (*25, 26*):

1) A search phase in which RNA polymerase repeatedly associates and disassociates with DNA until a promoter region is found.

2) An initial recognition step in which RNA polymerase forms a specific but relatively loose complex with the closed DNA helix at the promoter.

3) The "entry" of RNA polymerase to form an open complex in which four to six base pairs are opened. Binding of RNA polymerase to many promoters is very tight at this stage. Presumably polymerase also selects the appropriate DNA strand for use as a template at this point.

4) The open complex, in the presence of the four nucleoside triphosphates, is now able to initiate transcription at the start site.

5) If the entry site is separated from the start site, migration or "drift" of the RNA polymerase may be required between steps 3 and 4.

The properties of the sequence described in Fig. 6 suggest a model for how steps 3 through 5 might occur at the lac promoter (see Fig. 7).

A priori one would assume that the formation of the open complex would require both a specific sequence for unique recognition purposes and a sequence whose inherent transition temperature would allow denaturation stimulated by RNA polymerase (*26*). There is one region in the lac promoter that appears to fulfill both of these criteria. This is the A · T-rich region defined by promoter point mutations L241 and L305 and the repeating sequence CTTTA. We hypothesize that this is the RNA polymerase "entry site" which is essential for the formation of the open complex.

How does the CAP protein stimulate lac mRNA transcription? We propose that it does so by facilitating formation of the open complex at the entry site.

The A · T-rich entry site is bounded on both sides by sequences containing a large proportion of G · C base pairs (in Fig. 6 we show that 10 of 12 and 9 of 12 base pairs are G · C). We

Table 4. Computer analysis of sequence in block 2 (segments III, IV, and part of V). These oligonucleotides were used to construct the LAC sequence shown above by the procedures described in Table 3 and the text. Two solutions are possible representing the two possible orders of nucleotides 78 and 79. (See discussion in text on determination of the sequence, step 4.)

LAC T1 oligonucleotides	
T19	CUCACUCAUUAGG
T1	3GC
T7	CACCCCAGG
T17	CUUUACACUUUAUGC
T16	CUUCCGG
T10	CUCGU
T14A	UAUGU
T13	UUGU
T8	UGU
LAC ribonuclease A oligonucleotides	
P10A	AGGCA
CAL T1 and ribonuclease A oligonucleotides	
T19A	CCUAAUGA
T12	CCUGG
T15	UAAAGC
T16	CAUAAAGU
P11	GAGUG
P18	GGGGUG
P14	GGAAGCA
T3	CCGG
3′ end T26	UACGA

LAC sequence

CUCACUCAUUAGGCACCCCAG
30 40
GCUUUACACUUUAUGCUUCCGGCUCG
50 60 70
UAUGUUGUG
80

Table 5. Computer analysis of sequence in block 3 (segments V, VI, and part of IV). Solving the sequence for this block requires the sequence for CAL T26. To solve T26 we gave the ribonuclease A digestion products for T26, Table 1, to the computer sequencing program along with two independently derived sequences which are complementary to T26 and which put constraints on its sequence. The two sequences are GGAAUGU, an overlap of LAC P17 and T15, and a ribonuclease A partial digestion product UCGUAUGUUGUGU, in which the underlined bases are arranged arbitrarily. These two sequences must complement T26 because they do not complement any of the known oligonucleotides in block 3 of the CAL strand. The resulting sequence for T26 is shown above. Four sequences fit the data for block 3. Sequences 1 and 2 are not correct because they predict that deletion S20 would always delete T26 and only occasionally delete T20 and T9 (UG[A]). Our experiments give just the opposite result; T20 and T9 are always deleted by S20 while T26 is deleted only occasionally. Our data are insufficient to make a choice between sequences 3 and 4. We think sequence 4 is correct since it agrees with the operator sequence of Gilbert and Maxam (*11*).

CAL T1 oligonucleotides	
T10	CUG(U)
T23	UUUCCUGU
T9	UGU
T9	UGA
T20	AAAUUGU
T21	UUAUCCGC
T26	CUCACAAUUCCACACAA CAUACGA
CAL ribonuclease A oligonucleotides	
P10B	GAAAU(U)
LAC ribonuclease A and T1 oligonucleotides	
T15	AAUUGU
Part of T20	AACA
P11	GAGCG
P14	AGGAAACA

Possible CAL sequences

120 110 100
1 CUGUUUCCUGUUAUCCGCUCACAAUUCC
2 CUGUUUCCUGUGUUAUCCGCUCACAAUU
3 CUGUUUCCUGUGAAAUUGUGUUAUCCGC
4 CUGUUUCCUGUGUGAAAUUGUUAUCCGC
T9 T20

90 80 70
UUCCACACAACAUACGAAAUUGUGUGA
AAUUCCACACAACAUACGAAAUUGUGA
CCGCUCACAAUUCCACACAACAUACGA
CCGCUCACAAUUCCACACAACAUACGA
T26
deletion S20

assume that these sequences that contain many G · C pairs would tend to raise the transition temperature of the A · T-rich entry site, thus preventing formation of an open complex. Support for this assumption comes from the recent results of Burd, Wartell, and Wells (*27*), who studied the properties of a model double-stranded polymer $d(C_{15}A_{15}) \cdot d(T_{15}G_{15})$. The G · C sequences in this polymer have a strong stabilizing effect on the A · T sequences.

Binding of the CAP protein to its interaction site could destabilize the G · C-rich region next to it. This would lower the transition temperature of the entry site and allow formation of the open complex. In this model the G · C-rich region is a transducer, transmitting the effect of CAP binding to the entry site some 14 base pairs away. The function of the G · C "transducer" in the promoter is to keep the A · T-rich entry site closed in the absence of CAP binding, thereby allowing for catabolite repression (*27*).

This model suggests that it should be possible to activate the entry of RNA polymerase at the lac promoter by genetically altering the base composition of the G · C-rich blocks. A set of promoter mutations (class III) partially relieves the requirement for CAP (*18*) and may act in this manner. The only class III mutation which we have sequenced is p^rla, and it is a G · C to A · T transversion within a G · C-rich block (see Fig. 6).

After the open complex at the entry site is formed, RNA polymerase initiates transcription at the start site. The lac mRNA start site as determined in vitro by Maizels (*10*) and by Majors (*22*) is indicated in Figs. 5 and 6. The start site is not coincident with the proposed entry site, but rather is some 35 base pairs away.

The RNA polymerase binds tightly to DNA at the start site and can protect 40 to 45 base pairs from deoxyribonuclease digestion (*24, 28, 29*). In the lac UV5 promoter (*28*), in an fd promoter (*24*), and in the T7 A3 promoter (*29*) the start site is located near the center of the protected fragment. Thus the proposed lac entry site is probably not included in the protected fragment.

Since the lac entry site is not coincident with the start site and is not protected by RNA polymerase, we must assume some movement to the start site after entry. This movement could be "drift" (*25*) or a large conformational change in the enzyme, but in either case a translation of the opened base pairs is postulated.

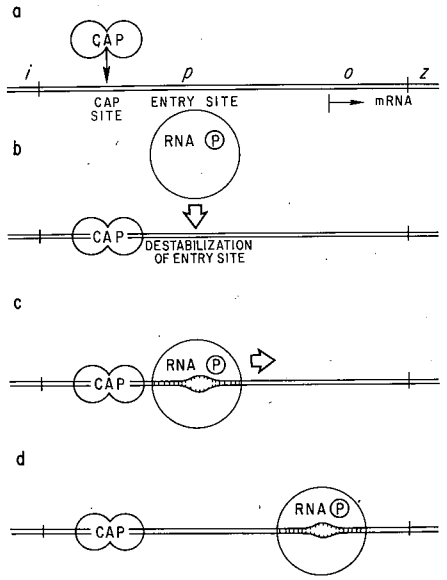

Fig. 7. A model for the initiation of lac transcription. The diagram is drawn to scale with RNA polymerase [RNA(P)], shown as a sphere 31 nucleotide pairs in diameter and CAP shown as two spheres each 11 nucleotide pairs in diameter. The CAP site is located at the region of symmetry notated in Figs. 5 and 6. The entry site is positioned at the high A · T block shown in Fig. 6 and the start site is at the 5′ end of lac mRNA (Fig. 5). (a) CAP binds to the CAP site. (b) CAP destabilizes the entry site thereby facilitating RNA(P) entry. (c) RNA(P) in the entry site "drifts" to the start site. (d) RNA(P) at the start site.

The model we have proposed assumes that the recognition site and the entry site are coincident. In this model CAP binding affects entry of RNA polymerase. Alternatively, the recognition site and the entry site could be separated and CAP could affect recognition alone (step 2 above). For example, CAP binding could cause a transition from the B to A form of DNA allowing RNA polymerase recognition at the site defined by the class II mutations. Also CAP could affect recognition by physically interacting with RNA polymerase. These models could allow the entry site and the start site to be coincident, thus avoiding the necessity of a translation of the opened base pairs.

Further sequence analysis of promoter mutations is in progress and we hope that this will help to define the promoter more clearly. It would be helpful to study the effect of the various promoter mutants in defined in vitro systems. For example, do class II promoter mutants affect formation of the closed complex (binding) or the open complex (entry)? It will be interesting to compare the nucleotide sequence of the lac control region with the sequences of other CAP regulated systems such as the arabinose operon. The comparisons may emphasize the important features of CAP action.

The DNA sequence defines the geometry of a control region but it gives only clues as to the mechanism of protein recognition of these sequences. A detailed description of control will require knowledge of the structures of the proteins involved, and more definite knowledge about the biochemical steps affected by promoter mutations.

Summary

The nucleotide sequence of the lac promoter-operator region has been determined. The 122 base pairs comprising this region include the recognition sites for RNA polymerase, the positive regulatory protein, CAP, and the negative regulatory protein, the repressor. Identification of mutant variants of the sequence combined with the in vitro biochemical studies of others has allowed us to tentatively identify the recognition site for each of these proteins, and to suggest how CAP might act at a distance to affect the interaction of RNA polymerase with the promoter.

References and Notes

1. F. Jacob and J. Monod, *J. Mol. Biol.* **3**, 318 (1961).
2. F. Jacob, A. Ullman, J. Monod, *C. R. Hebd. Acad. Sci. Paris* **258**, 3125 (1964); J. Scaife and J. Beckwith, *Cold Spring Harbor Symp. Quant. Biol.* **31**, 403 (1966); W. S. Reznikoff, *Annu. Rev. Genet.* **6**, 133 (1972).
3. W. Gilbert and B. Muller-Hill, *Proc. Natl. Acad. Sci. U.S.A.* **58**, 2415 (1967); A. D. Riggs, R. R. Newby, S. Bourgeois, M. Cohn, *J. Mol. Biol.* **34**, 365 (1968).
4. R. S. Makman and E. W. Sutherland, *J. Biol. Chem.* **240**, 1309 (1965).
5. B. de Crombrugghe, B. Chen, W. Anderson, P. Nisley, M. Gottesman, I. Pastan, R. Perlman, *Nat. New Biol.* **231**, 139 (1971); L. Eron and R. Block, *Proc. Natl. Acad. Sci. U.S.A.* **68**, 1828 (1971).
6. W. M. Barnes, W. S. Reznikoff, F. Blattner, R. C. Dickson, J. Abelson, in preparation.
7. W. M. Barnes, R. B. Siegel, W. S. Reznikoff, *Mol. Gen. Genet.* **129**, 201 (1974); W. M. Barnes, thesis, University of Wisconsin (1974).
8. F. Sanger, G. G. Brownlee, B. G. Barrell, *J. Mol. Biol.* **13**, 373 (1965); G. G. Brownlee, *Determination of Sequences in RNA* (Elsevier, New York, 1972).
9. Written by William Atkinson of San Diego.
10. N. Maizels, *Proc. Natl. Acad. Sci. U.S.A.* **70**, 3585 (1973).
11. W. Gilbert and A. Maxam, *ibid.*, p. 3581.
12. K. Beyreuther, K. Adler, H. Geisler, A. Klemm, *ibid.*, p. 3576.
13. P. Lebowitz, S. M. Weissman, C. M. Radding, *J. Biol. Chem.* **246**, 5120 (1971); G. Pieczenik, B. Barrell, M. Gefter, *Arch. Biochem. Biophys.* **152**, 152 (1972); J. Dahlberg and F. Blattner, in *Virus Research*, C. F. Fox and W. S. Robinson, Eds. (Academic Press, New York, 1973), p. 533; T. Ikemura and J. E. Dahlberg, *J. Biol. Chem.* **248**, 5024 (1973).

14. W. S. Reznikoff, C. A. Michels, T. G. Cooper, A. E. Silverstone, B. Magasanik, *J. Bacteriol.* **117**, 1231 (1974); D. Mitchell, W. S. Reznikoff, J. Beckwith, in press.
15. W. S. Reznikoff, R. B. Winter, C. K. Hurley, *Proc. Natl. Acad. Sci. U.S.A.* **71**, 2314 (1974).
16. A. Gierer, *Nature (Lond.)* **212**, 1480 (1966).
17. J. C. Wang, M. D. Barkley, S. Bourgeois, *ibid.* **251**, 247 (1974).
18. J. H. Miller, K. Ippen, J. G. Scaife, J. R. Beckwith, *J. Mol. Biol.* **38**, 413 (1968); R. Arditti, T. Grodzicker, J. Beckwith, *J. Bacteriol.* **114**, 652 (1973); J. D. Hopkins, *J. Mol. Biol.* **87**, 715 (1974).
19. J. Beckwith, T. Grodzicker, R. R. Arditti, *J. Mol. Biol.* **69**, 155 (1972).
20. Deletion X8630 is one of several *tonB-lacI⁻ lac⁺* deletions whose isolation is described by J. H. Miller, W. S. Reznikoff, A. E. Silverstone, K. Ippen, E. R. Signer, J. R. Beckwith [*J. Bacteriol.* **104**, 1273 (1970)].
21. A. D. Riggs, G. Reiness, G. Zubay, *Proc. Natl. Acad. Sci. U.S.A.* **68**, 1222 (1971).
22. J. Majors, personal communication.
23. More extensive repeats are discernible in this interval if A·G and C·T degeneracies are allowed; that is, the LAC sequence purCTTTApyrpurCTTpyr is present as an overlapping repeat (pyr, pyrimidine; pur, purine).
24. H. Schäller, C. Gray, K. Herrmann, *Proc. Natl. Acad. Sci. U.S.A.*, in press.
25. W. Zillig, K. Zechel, D. Rabussay, M. Schackner, V. Sethi, P. Palm, A. Heil, W. Seifert, *Cold Spring Harbor Symp. Quant. Biol.* **35**, 47 (1970); F. R. Blattner, J. E. Dahlberg, J. K. Boettiger, M. Fiandt, W. Szybalski, *Nat. New Biol.* **237**, 232 (1972); J. Saucier and J. Wang, *ibid.* **239**, 167 (1972).
26. M. Chamberlin, *Annu. Rev. Biochem.* **43**, 721 (1974).
27. J. Burd, R. Wartell, R. D. Wells, *J. Biol. Chem.*, in press. Binding of actinomycin D to the model polymer $d(C_{15}A_{15}) \cdot d(T_{15}G_{15})$ raises the melting temperature of the A·T sequences. Burd *et al.* have termed the transmitted affect of binding "telestability" and have independently suggested that such interactions could explain the mode of action of CAP in stimulating lac expression. They also suggest that this effect might explain the phenotypes of certain lac promoter mutations.
28. J. Gralla, personal communication.
29. D. Pribnow, personal communication.
30. T. Pinkerton, G. Paddock, J. Abelson, *J. Biol. Chem.* **248**, 6348 (1973).
31. The exact location of the start site shows a slight ambiguity since occasionally the lac message starts with the preceding G (*10*, *22*).
32. We thank F. Blattner and J. Dahlberg for suggestions and materials; W. Atkinson for computer programming; R. Burgess for RNA polymerase; C. Hurley, P. Johnson, D. MacDonald, A. Otsuka, and K. Thornton for technical assistance; J. Beckwith, S. Bourgeois, T. Grodzicker, J. Hopkins, J. Miller, and D. Mitchell for bacterial strains; and J. Burd, W. Gilbert, J. Gralla, D. Pribnow, J. Majors, H. Schäller, and R. Wells for permission to quote their results prior to publication. This work was supported by NIH grants 1-R01-GM-19670 and CA 10984, by NSF grant GB-20462, by the Cancer Research Coordinating Committee of the University of California, and by the Wisconsin Alumni Research Foundation. J.A. was supported by a faculty research award from the American Cancer Society. W.S.R. was supported by a career development award 5-K04-GM-30970 from NIH and W.M.B. was supported by NIH training grant GM-00236.