Proc. Natl. Acad. Sci. USA
Vol. 79, pp. 2427–2431, April 1982
Biochemistry

Sequence of *galR* gene indicates a common evolutionary origin of *lac* and *gal* repressor in *Escherichia coli*

(DNA sequence/operator/protein–DNA recognition)

BRIGITTE VON WILCKEN-BERGMANN AND BENNO MÜLLER-HILL

Institut für Genetik der Universität zu Köln, Köln, Federal Republic of Germany

Communicated by Walter Gilbert, November 30, 1981

ABSTRACT **The nucleotide sequence of the *galR* gene of *Escherichia coli*, which codes for galactose repressor, has been determined. The subunits of *gal* repressor are predicted to consist of 343 residues, including the NH_2-terminal methionine. Twenty-six of the predicted NH_2-terminal 55 residues of *gal* repressor are identical to the NH_2-terminal residues of *lac* repressor. Additional homologies appear between residues 165 and 200, between residues 235 and 255, and around residue 325.**

The galactose operon of *Escherichia coli* is negatively controlled by a repressor, as is the *lac* operon (1). The *gal* operator region is well defined by sequence analysis of wild-type and mutant (o^c) DNA (2, 3). Recessive constitutive R^- (1) and negative dominant R^s (4) mutations correspond to I^- and I^s mutations in the *lac* system (5). The inducer molecules for the two repressors, galactose and allolactose, are chemically related. Galactose, the inducer of the *gal* repressor (6) also induces the *lac* operon (7), whereas methyl thiogalactoside, a gratuitous inducer for the *lac* operon (8), behaves as an anti-inducer with *gal* repressor (6). Thus the inducer-recognizing parts of the two repressors have similar specificities. The *galR* gene is not linked to the *gal* operon, but maps on the opposite side of the circular *E. coli* chromosome close to *lysA* (1).

We cloned the *galR* gene and determined its nucleotide sequence. We found the *gal* and *lac* repressors to be structurally similar, indicating a common origin in evolution.

MATERIALS AND METHODS

E. coli K-12 BMH71-18 $[lac\text{-}pro]_{del}$/F′ *pro*$^+$*lacI*^{q}*ZM15* has been described (9). LF461 and HS340 have the genotypes *galR*^{s} and $[galR\text{-}lysA]_{del}$, respectively (4). Both strains were kindly provided by P. Starlinger. Phage M13mp2 is described in ref. 10, M13mp2Bam in ref. 11, and M13mp7 in ref. 12. M13mp51 was kindly provided by B. Gronenborn. Plasmids pBR322 and pSP16 are described in refs. 13 and 14, respectively.

Restriction endonucleases *Bam*HI, *Hae* III, *Hha* I, *Hin*cII, *Hin*dIII, *Hin*fI, *Hpa* II, *Pst* I, *Sau* 3A, and *Taq* I were purchased from Bethesda Research Laboratories (Neu-Isenburg, Federal Republic of Germany), *Acc* I from Biolabs (Bad Schwalbach, Federal Republic of Germany), *Eco*RI and DNA polymerase I (large fragment) from Boehringer Mannheim. Phage T4 DNA ligase was purified according to ref. 15. *Eco*RI linker was obtained from Collaborative Research (Waltham, MA). The chemicals used for sequence analysis were from the same firms as in ref. 16. Eosin/methylene blue (EMB) galactose plates are described in ref. 17. Procedures for cloning and sequence analysis were the same as in ref. 16. Galactose kinase activity was measured as in ref. 18.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

RESULTS

Among the ampicillin-resistant colonies recovered after a "shotgun" experiment with *Hin*dIII fragments of *E. coli* chromosomal DNA ligated into plasmid pBR322, one clone yielded a plasmid that made *E. coli gal*$^+$ colonies look slightly galactose-negative on EMB galactose indicator plates. However, growth on galactose was normal. To see whether this effect was due to an overproduction of *gal* repressor we transferred this plasmid (pRK3) into a *gal*$^+$ strain (HS340) carrying a *galR-lysA* deletion. The plasmid conferred lysine independence to this strain and repressed the level of galactokinase strongly (data not shown). The overproduction of *gal* repressor by plasmid pRK3 was sufficient to make strain LF461 (*galR*^{s}) grow on galactose. A similar change in dominance has been reported previously for the I^q mutation in the case of the *lac* repressor. I^q, which causes 10-fold overproduction of *lac* repressor, is dominant over the non-inducible I^s gene because of subunit mixing (19).

Fig. 1 gives a physical map of pRK3 and a summary of the construction of two pairs of hybrid M13 phages that carry different parts of the *galR* coding region in both orientations. M13mp51n4 and M13mp51n10—referred to as n4 and n10, respectively—lack 31 NH_2-terminal codons, whereas M13mp2/1013 and M13mp2/2213—referred to as 1013 and 2213, respectively—lack 14 COOH-terminal codons. A detailed description of the construction and phenotypes of these and other plasmids that overproduce active *gal* repressor will be given elsewhere.

We used the single-stranded DNA of the four hybrid phages n4, n10, 1013, and 2213 as templates for sequence analysis according to the method of Sanger *et al.* (20). Small restriction fragments required to prime polymerase action were prepared from double-stranded replicative form molecules of the same phage. They also served as a source for fragments whose sequences were determined according to the method of Maxam and Gilbert (21). The way in which we obtained the nucleotide sequence of the *galR* gene from both strands is outlined in Fig. 2. The DNA sequence of the *galR* gene is presented together with the sequence of the *lacI* gene in Fig. 3.

DISCUSSION

The protein sequence deduced from the one open reading frame of the *galR* gene (Fig. 3) resembles the protein sequence of *lac* repressor (22, 23). There is only one AUG start codon, which is preceded by a Shine–Dalgarno sequence (24) T-A-A-G-G at a distance of seven base pairs. K. Beyreuther and R. Ehring (personal communication) have used *galR*-carrying plasmid DNA to synthesize *gal* repressor *in vitro* in the presence of labeled methionine and alanine. By automated sequence analysis they have shown that *gal* repressor indeed begins with methi-

Abbreviations: EMB, eosin/methylene blue; bp, base pair(s).

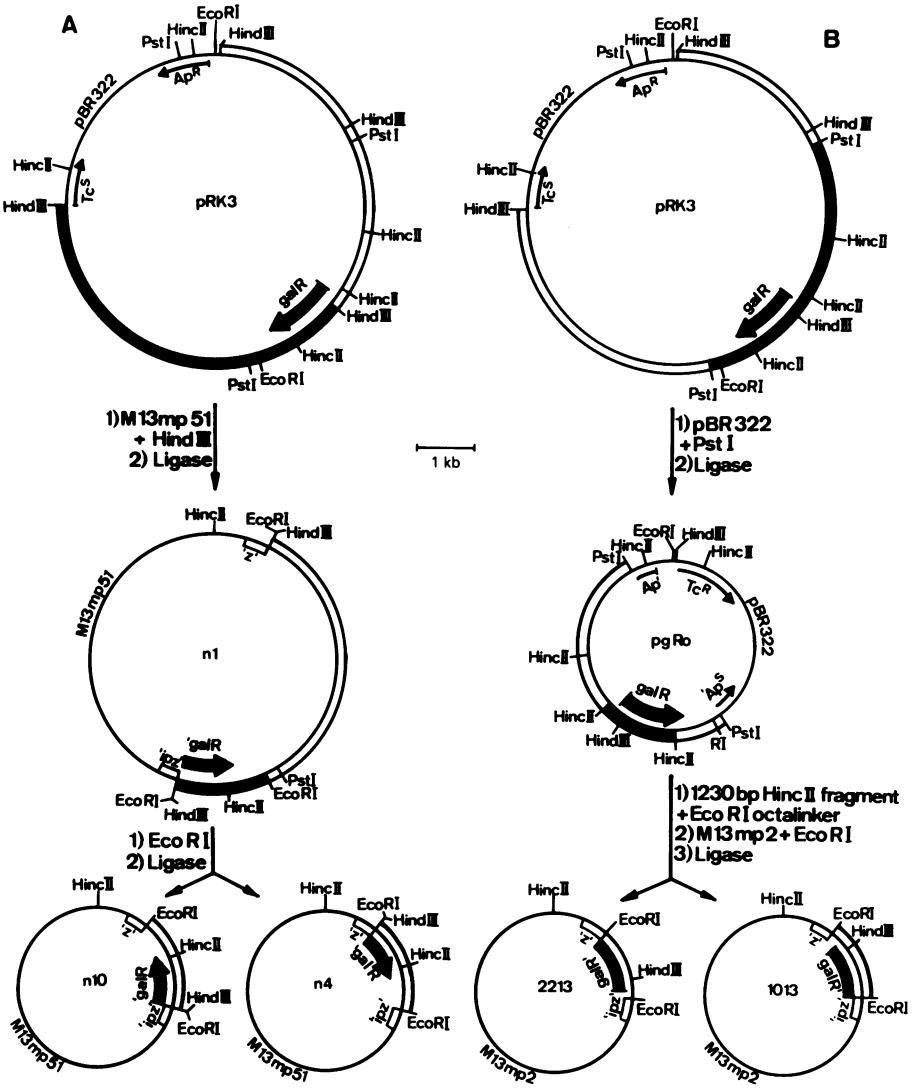

FIG. 1. (*A*) Construction of hybrid M13 phages containing parts of the *galR* gene. Purified pRK3 DNA was digested with *Hin*dIII. The largest fragment was inserted into the *Hin*dIII site of M13mp51. Purified replicative form II molecules of such phage were digested with *Eco*RI. The resulting mixture of fragments was incubated with T4 DNA ligase and used to transform strain BMH71-18. The size and the orientation of the inserts was determined by digestion with *Eco*RI or by digestion with *Hin*cII and *Hin*dIII. (*B*) Purified pRK3 DNA was digested with *Pst* I. Fragments from the resulting mixture were inserted into the *Pst* I site of pBR322. *E. coli* HS340 was transformed. Pink colonies were purified from EMB galactose tetracycline-containing plates. Restriction analysis showed that plasmid pgRO carried a 4.8-kilobase (kb) fragment in the *Pst* I site of pBR322. From purified pgRO plasmid DNA a *Hin*cII fragment of 1,230 base pairs (bp) was isolated. *Eco*RI octanucleotide linkers were ligated to its end with T4 DNA ligase. After treatment with *Eco*RI the fragment was inserted into the *Eco*RI site of M13mp2 (10). Strain BMH71-18 was transformed with the ligation mixture. Plasmid 1013 was isolated from a white plaque; plasmid 2213, from a light blue plaque on the proper indicator plates (9, 10). To determine the presence and the orientation of the inserts, equal amounts of single-stranded 1013 and 2213 phage DNA were hybridized with n4 and n10 phage DNA. Hybrid molecules were detected on agarose gels as slower-moving bands. All plasmid and phage DNAs are drawn to scale. Inserted *E. coli* DNA is represented by a double line and subcloned DNA, by a black line. The position and direction of transcription of the *galR* gene are indicated by an arrow. Ap^R, ampicillin resistance; Tc^S, tetracycline sensitivity; p, promoter; *i* and *z*, structural genes for *lac* repressor and β-galactosidase, respectively. Single quotation marks before or after gene symbols indicate incomplete genes.

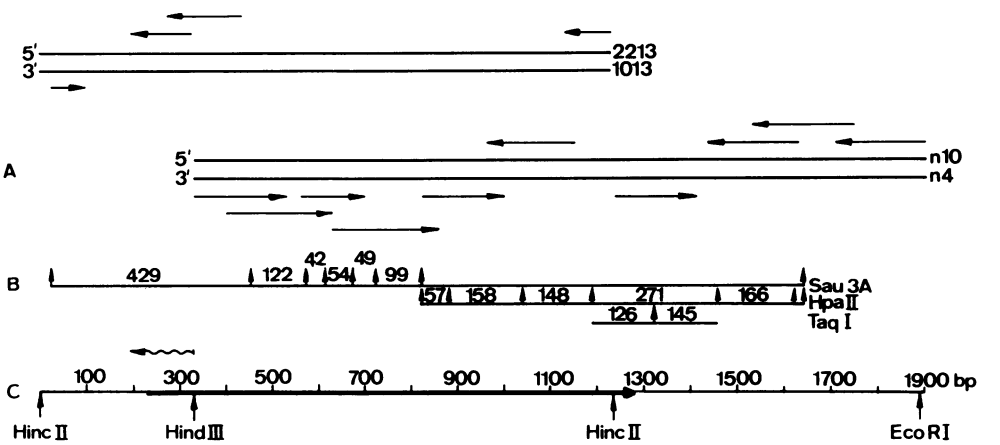

FIG. 2. Partial restriction map of the *galR* gene. (*A*) Size of the *galR* fragments carried in the single-stranded phages 2213, 1013, n4, and n10, which were used as templates with the dideoxy chain termination method (20). Sequence data are indicated as arrows above or below the corresponding templates. The primers are not shown. They were generated by digestion of the cloned fragments with *Hae* III, *Hha* I, *Hpa* II, *Hin*fI, *Sau*3A, and *Taq* I, either singly or in various combinations. (*B*) In addition to the templates and sequences shown in *A*, we prepared subclones of all the *Sau*3A fragments in both orientations in M13mp2Bam (11). From the 810-bp *Sau*3A fragments we further subcloned the *Hpa* II and *Taq* I fragments in the *Acc* I site of M13mp7 (12). Using the 26-bp *Eco*RI fragment from pSP16 (14) as a primer with these templates, we received the complete nucleotide sequence of the *galR* gene from both strands except for the NH_2-terminal region. The size and position of the subcloned fragments are drawn to scale. (*C*) Physical map of the region whose sequence was determined; the location of the *galR* gene is indicated by an arrow. The wavy line above the map represents sequence data obtained with the chemical modification method of Maxam and Gilbert (21) after 3' labeling of the *Hin*dIII site of the fragment sequenced.

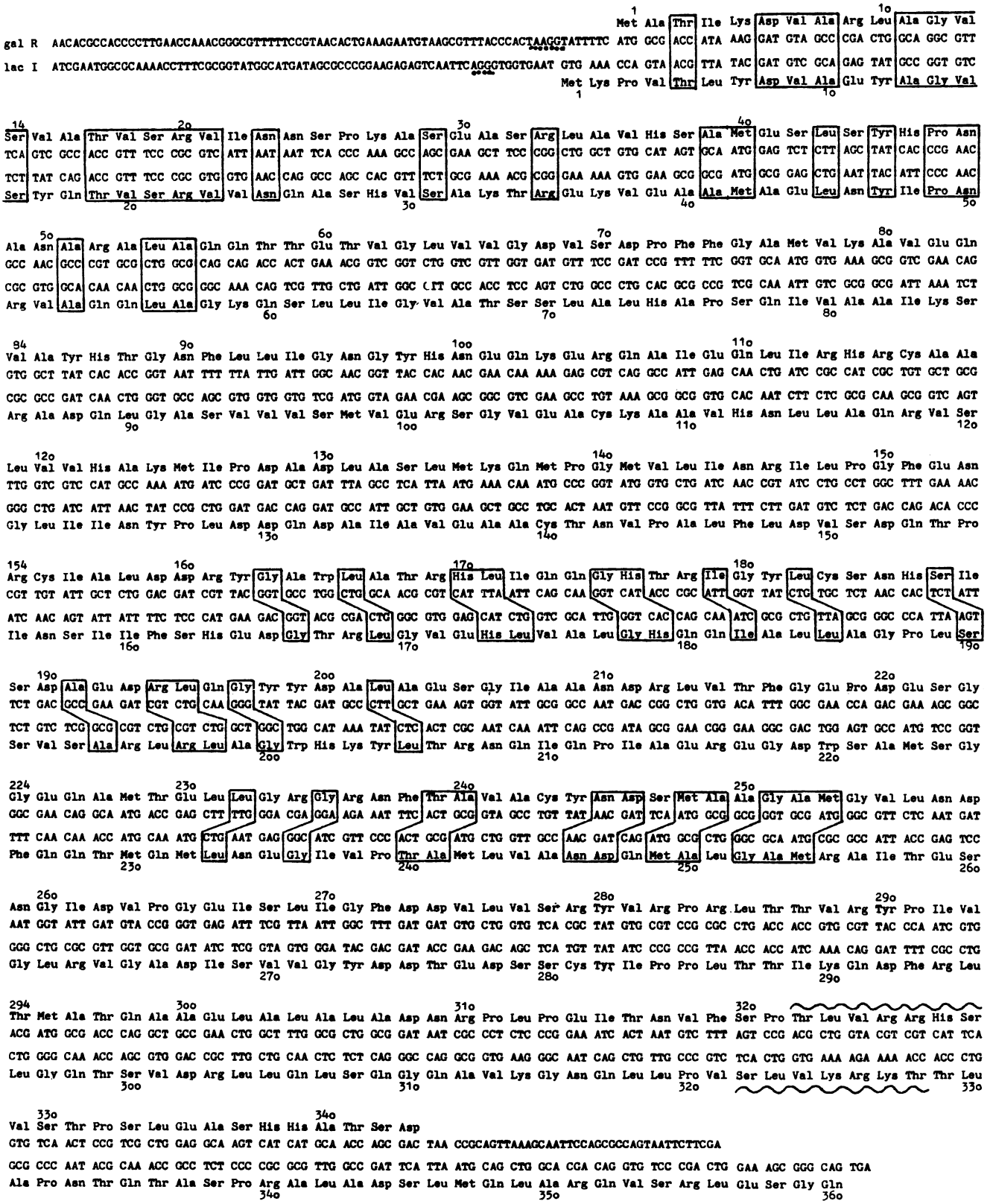

FIG. 3. The DNA sequence of the *galR* gene and its deduced protein sequence (upper lines) are compared to the sequences of *lac* repressor (22, 23). The sequences have been aligned in a manner to maximize the homology of the NH_2-terminal regions. Identical amino acids in regions of homology are shown in boxes. The COOH-terminal basic sequences are indicated by wavy lines. The assumed ribosomal binding sites (24) are marked by dotted lines.

onine and contains alanine in positions 2, 8, and 11, as the DNA sequence predicts.

Four regions of homology between *gal* and *lac* repressor can be found (Fig. 3): between residues 1 and 55, the headpiece of *lac* repressor, 26 amino acids are identical; between residues 165 and 200, 23 out of 35 amino acids are identical; and between residues 235 and 255, 10 out of 20 amino acids are identical; finally, the basic sequence that is found in *lac* repressor between

P22CII	45	Pro	Asn	Gly	Glu	Asn	Leu	Leu	Ala	Leu	Ser	Lys	Ala	Leu	Gln	Cys	Ser	60
λCI	57	Leu	Asn	Ala	Tyr	Asn	Ala	Ala	Leu	Leu	Ala	Lys	Ile	Leu	Lys	Val	Ser	72
galR	44	Ser	Tyr	His	Pro	Asn	Ala	Asn	Ala	Arg	Ala	Leu	Ala	Gln	Gln	Thr	Thr	59
lacI	46	Asn	Tyr	Ile	Pro	Asn	Arg	Val	Ala	Gln	Gln	Leu	Ala	Gly	Lys	Gln	Ser	61

FIG. 4. Amino acid sequences of the NH_2-terminal regions of phage P22 repressor, λ repressor (37), *gal* repressor, and *lac* repressor (22). Identical amino acids are shown in boxes.

residues 322 and 328 appears in *gal* repressor at the same position. The conservation of the sequence homologies from residue 165 to 255 may reflect the similarity of the inducer-binding sites, which bind rather similar molecules, galactose and alkyl-galactosides. The region between residues 155 and 270 is involved in inducer binding in *lac* repressor according to the map positions of i^s mutations (25, 26). We think that a similar secondary structure is conserved, too, in both repressors. For example, some prolines in one sequence have been exchanged for asparagines in the other, indicating β bends. This is even true in regions where no sequence homology can be detected. Thus Asn-142, Pro-188, Pro-212, and Pro-239 of *lac* repressor correspond to proline or asparagine, respectively, in *gal* repressor. The conservation of protein sequences in various regions of *lac* and *gal* repressor is consistent with the assumption that a precursor gene of the *lac* and *gal* repressor genes has been duplicated. The map positions of *galR* and *lacI* are just opposite on the circular chromosome of *E. coli* (1). This could result from a doubling of the *E. coli* chromosome as proposed by Riley and Anilionis (27). The almost complete absence of DNA homology between the two genes reflects the long time passed since the duplication happened.

We now would like to discuss in some detail the implications of the homology found in the NH_2-terminal region of the two repressors. The clustering of all fully constitutive I^{-d} mutations between codon 3 and 58 of *lac* repressor has led to the proposal that this region exclusively recognizes and binds *lac* operator (25, 28, 29). Some *in vitro* experiments are also consistent with this proposal. The NH_2-terminus of *lac* repressor is a stable domain (30), which can be proteolytically cleaved intact from *lac* repressor. This so-called long headpiece of residues 1–59 protects *lac* operator against methylation in a manner similar to that of *lac* repressor (31).

Evidence that the NH_2-terminal region of *gal* repressor is necessary for *gal* operator binding is provided by subclone n10 (Fig. 1), in which five NH_2-terminal codons of the *lacZ* gene (10) are joined in phase to codon 31 of the *galR* gene. Subclone n10 changes the phenotype of a $galR^s$ strain from galactose-negative to galactose-utilizing, and it does not repress galactokinase production in a $galR^-$ strain (data not shown). This resembles the behavior of $lacI^{-d}$ mutations in which the NH_2-terminus of *lac* repressor has amino acid exchanges or is replaced by another NH_2-terminus (25). Subclone n4, which contains the same segment of DNA as n10 but in the opposite orientation, does not change the phenotype of $galR^s$; thus dominance cannot be caused by *gal* operator-like sequences present on the phage DNA.

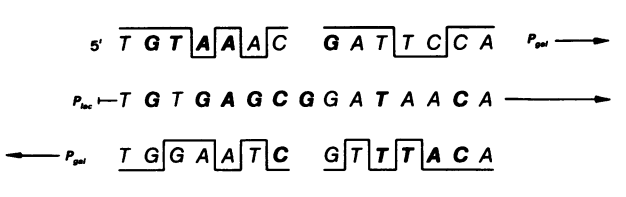

FIG. 5. The *gal* operator is compared to the *lac* operator in both orientations in order to visualize differences in homology due to asymmetry of the sequences. All sequences are written in 5′ to 3′ direction as indicated in the upper line. The sequence of one strand of the *lac* operator is given in the middle (39). Above and below the two strands of the *gal* operator (3) are aligned in opposite orientations. P_{gal} and P_{lac} stand for the *gal* and *lac* promoters, respectively, and the arrows indicate the direction of transcription. The gap corresponds to the symmetry center. The homology is stressed by a line that excludes differing nucleotides. The positions where base pair exchanges have been found to result in o^c mutations (3, 39) are displayed in boldface characters.

Matthews proposed a different model of *lac* repressor–operator interaction (32). She proposed that sequences in the core of *lac* repressor recognize operator and that the NH_2-terminal headpiece binds nonspecifically to DNA. The major argument for her model is that the tryptic core of *lac* repressor recognizes specifically *lac* operator without binding nonspecifically to DNA (33). She assumes the core she used for her experiments begins with residue 60. The amino acid analysis of the *lac* repressor core she used indicates, however, that it begins predominantly with residue 52 [ref. 33; see also ref. 34, in which the NH_2-terminal residues of the core are said to be glutamine (i.e., residue 60) and valine (i.e., residue 52)]. Five different strong constitutive I^{-d} mutations were localized by protein sequence analysis in the region between residue 53 and 58 (29, 35). Thus this region is partly responsible for operator recognition (25). In summary, we think that the results of Matthews are quite compatible with the model proposed by us.

K. Beyreuther and M. Cossmann (personal communication) have found the Lys-59 residues of two subunits of *lac* repressor to be crosslinkable by suberimidate. This leads Beyreuther to suggest that the two headpieces recognize the region around the symmetry center of the operator, using sequences 51 to 59. Analysis of the secondary structure of *gal* and *lac* repressors according to Chou and Fasman (36) leads us to predict that the region between residue 50 and residue 60 is α-helical. This leads us to speculate that α-helical arms from two subunits approach the symmetry center of operator with Lys-59 and follow the deep grooves away from the symmetry center until the helix is broken by the Pro-Asn sequence. There exists also sequence homology between the NH_2-terminal sequences of the repressors of phage λ, phage P22 (37), and this region between residues 48 and 61 of *lac* and *gal* repressor (Fig. 4). The sequence homology between the two phage repressors is the only one in the NH_2-terminal part of the repressor sequences (37). It should be recalled that it is the NH_2-terminal part of λ repressor that recognizes λ operator (38). The homology may point to convergent evolution of a similar type of protein–DNA recognition through α-helical arms in all four repressors.

The close similarity of the NH_2-termini of *gal* and *lac* repressors should be reflected in the structure of the *gal* (3) and *lac* operators (39). A comparison of the two operators (Fig. 5) shows indeed a close similarity. The central G·C pair of *lac* operator is lacking from the *gal* operator, but five or four out of seven base pairs on each side are identical.

Note Added in Proof. Recently the following mode of repressor–operator recognition occurred to us: Residues 53 to 64 of two subunits of *lac* re-

pressor could form an antiparallel β-sheet in the minor groove of *lac* operator with Gly-58 placed over the G·C pair in the symmetry center. The other G·C pairs would then be recognized by Gln and Leu.

We thank K. Otto for technical help, B. Gronenborn, R. Klaer, and P. Starlinger for providing strains, K. Beyreuther and R. Ehring for the communication of unpublished results, and the Deutsche Forschungsgemeinschaft for support through SFB 74.

1. Buttin, G. (1963) *J. Mol. Biol.* **7**, 183–205.
2. Musso, R., Di Lauro, R., Rosenberg, M. & de Crombrugghe, B. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 106–110.
3. Adhya, S. & Miller, W. (1979) *Nature (London)* **279**, 492–494.
4. Saedler, H., Gullon, A., Fiethen, L. & Starlinger, P. (1968) *Mol. Gen. Genet.* **102**, 79–88.
5. Willson, C., Perrin, D., Cohn, M., Jacob, F. & Monod, J. (1964) *J. Mol. Biol.* **8**, 582–594.
6. Buttin, G. (1963) *J. Mol. Biol.* **7**, 164–182.
7. Riggs, A. D., Newby, R. F. & Bourgeois, S. (1970) *J. Mol. Biol.* **51**, 303–314.
8. Jacob, F. & Monod, J. (1961) *J. Mol. Biol.* **3**, 318–356.
9. Messing, J., Gronenborn, B., Müller-Hill, B. & Hofschneider, P.-H. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 3642–3646.
10. Gronenborn, B. & Messing, J. (1978) *Nature (London)* **272**, 375–377.
11. Rothstein, R. J., Lau, L. F., Bahl, C. P., Narang, S. A. & Wu, R. (1979) *Methods Enzymol.* **68**, 98–109.
12. Messing, J., Crea, R. & Seeburg, P. H. (1981) *Nucleic Acids Res.* **9**, 309–321.
13. Bolivar, F., Rodriguez, R. L., Greene, P. J., Betlach, M. C., Heynecker, H. L., Boyer, H. W., Cross, J. H. & Falkow, S. (1977) *Gene* **2**, 95–113.
14. Anderson, S., Gait, M. J., Mayol, L. & Young, I. G. (1980) *Nucleic Acids Res.* **8**, 1731–1745.
15. Tait, R. C., Rodriguez, R. L. & West, R. W., Jr. (1980) *J. Biol. Chem.* **255**, 813–815.
16. Büchel, D., Gronenborn, B. & Müller-Hill, B. (1980) *Nature (London)* **283**, 541–543.
17. Miller, J. H. (1972) *Experiments in Molecular Genetics* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY).
18. Wilson, D. B. & Hogness, D. S. (1966) *Methods Enzymol.* **8**, 229–240.
19. Müller-Hill, B., Crapo, L. & Gilbert, W. (1968) *Proc. Natl. Acad. Sci. USA* **59**, 1259–1264.
20. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
21. Maxam, A. & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 560–564.
22. Beyreuther, K., Adler, K., Geisler, N. & Klemm, A. (1973) *Proc. Natl. Acad. Sci. USA* **70**, 3576–3580.
23. Farabough, P. (1978) *Nature (London)* **274**, 765–769.
24. Shine, J. & Dalgarno, L. (1974) *Proc. Natl. Acad. Sci. USA* **71**, 1342–1346.
25. Müller-Hill, B. (1975) *Prog. Biophys. Mol. Biol.* **30**, 227–252.
26. Miller, J. (1978) in *The Operon*, eds. Miller, J. H. & Reznikoff, W. S. (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY), pp. 31–88.
27. Riley, M. & Anilionis, A. (1978) *Annu. Rev. Microbiol.* **32**, 519–560.
28. Adler, K., Beyreuther, K., Fanning, E., Geisler, N., Gronenborn, B., Klemm, A., Müller-Hill, B., Pfahl, M. & Schmitz, A. (1972) *Nature (London)* **237**, 322–327.
29. Beyreuther, K. (1978) in *The Operon*, eds. Miller, J. H. & Reznikoff, W. S. (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY), pp. 123–154.
30. Buck, F., Rüterjans, H. & Beyreuther, K. (1978) *FEBS Lett.* **96**, 335–338.
31. Ogata, R. T. & Gilbert, W. (1978) *Proc. Natl. Acad. Sci. USA* **75**, 5851–5854.
32. Dunaway, M., Manley, S. P. & Matthews, K. S. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 7181–7185.
33. O'Gorman, R. B., Dunaway, M. & Matthews, K. S. (1980) *J. Biol. Chem.* **255**, 10100–10106.
34. Dunaway, M. & Matthews, K. S. (1980) *J. Biol. Chem.* **255**, 10120–10127.
35. Weber, K., Files, J. G., Platt, D., Ganem, D., Miller, J. H. (1975) *Proc. Natl. Acad. Sci. USA* **69**, 3624–3628.
36. Chou, P. Y. & Fasman, G. D. (1974) *Biochemistry* **13**, 223–245.
37. Sauer, R. T., Pan, J., Hopper, P., Hehir, K., Brown, J. & Poteete, R. (1981) *Biochemistry* **20**, 3591–3598.
38. Sauer, R. T., Pabo, C. O., Meyer, B. J., Ptashne, M. & Backmann, K. C. (1979) *Nature (London)* **279**, 396–400.
39. Gilbert, W., Majors, J. & Maxam, A. (1976) in *Organisation and Expression of Chromosomes*, ed. Allfrey, V. G. (Dahlem Konferenzen, Berlin), pp. 167–178.