A Technique for Expressing Eukaryotic Genes in Bacteria

Leonard Guarente, Thomas M. Roberts, Mark Ptashne

Summary. Methods are described that allow efficient expression in *Escherichia coli* of cloned eukaryotic genes. The methods require that the coding sequence of the gene in question be available in a form uninterrupted by intervening sequences (for example, as a complementary DNA clone). The gene products are synthesized unfused to other amino acid sequences. The genetic manipulations are simple, and require the plasmids described and commercially available enzymes.

There are two salient differences between the genetic signals required for gene expression in prokaryotes and eukaryotes. The first difference is in those sequences that direct RNA polymerase binding to DNA and initiation of transcription, that is, promoters. Although the exact nature of eukaryotic promoters is not understood, it is apparent that these signals do not, in general, function in bacteria. The second difference is in the signals that allow messenger RNA (mRNA) to be translated into protein. In eukaryotic mRNA's, the only sequence required for eukaryotic ribosome binding and translation initiation seems to be the AUG (A, adenine; U, uracil; G, guanine) encoding the NH_2-terminal methionine which, in general, is the first AUG triplet of the message (*1*). Also required in at least some cases is a post-transcriptional modification (that is, cap) at the immediate 5′ end of the transcript [for reviews, see (*1*)]. In bacterial mRNA's, there is a sequence in addition to the AUG that is apparently required for efficient bacterial ribosome binding and initiation of translation (*2, 3*). This 3- to 12-base pair (bp) sequence, known as the Shine-Dalgarno (SD) sequence, occurs in the 5′ untranslated leader region of the mRNA; the SD sequence begins 3 to 11 nucleotides upstream from the AUG. This SD sequence is complementary to the 3′ end of 16*S* ribosomal RNA, and the duplexing allowed by this complementarity is thought to play a role in stabilizing the initiation complex formed between the mRNA and the ribosome (*3*). It is, therefore, not surprising that mRNA's which lack SD sequences are not efficiently read by the *Escherichia coli* translational machinery.

The above considerations suggest that the minimal requirements for the expression in *E. coli* of a cloned eukaryotic gene are that it be transcribed from a bacterial promoter and that the resultant mRNA bear an appropriately positioned SD sequence. Our method satisfies these requirements of fusing a fragment of DNA encoding a promoter and leader transcript of an *E. coli* gene is fused to a second fragment of DNA that bears the coding sequences of the eukaryotic gene. The mRNA produced from this fusion contains a "hybrid" ribosome binding site (*4*) consisting of the SD sequence from the *E. coli* gene and the initiating AUG of the eukaryotic gene. As a convenient source of a promoter and leader, we have used a DNA fragment of the *lac* operon encoding the promoter and the *lacZ* leader through the SD sequence. This fragment ends two base pairs before the initiating ATG (T, thymine) codon of *lacZ*. The promoter fragment bears a mutation (UV^5) that renders it functional in the absence of the CAP protein (catabolite gene action protein) and cyclic adenosine monophosphate ordinarily required to stimulate this promoter. Moreover, the fragment bears the *lac* operator, the site at which the *lac* repressor controls the promoter. As a consequence, the levels of protein synthesized under direction of this promoter can be regulated by inducers of the *lac* operon such as isopropyl thiogalactoside (*5*).

The placement of the promoter fragment and, hence, the *lacZ* SD sequence relative to the ATG at the start of the eukaryotic gene is critical. The techniques we have used for this purpose have progressed through several stages. The promoter fragment was first used to express the phage λ*cI* gene that encodes the λ repressor (*4*). The placement of the promoter in that case depended upon the availability of a restriction enzyme site at a specific location near the start of *cI* (*6*). The method was then generalized (with the use of λ's *cro* gene as a model system) to include genes that do not have restriction sites located very close to their starts (*7, 8*).

We begin by cloning the gene of interest on a plasmid (the coding sequences of eukaryotic—but not of prokaryotic—genes are often interrupted by intervening sequences and therefore the eukaryotic gene is usually cloned as a complementary DNA copy of the corresponding mRNA). We then introduce, if one does not already exist, a restriction enzyme site in the 5′ flanking region within approximately 100 bp of the gene start. For this purpose, we usually use a synthetic linker fragment (*9*) that encodes a restriction site not found elsewhere on the plasmid. Next, we open the plasmid at that site and excise varying amounts of DNA with exonuclease. We then insert the promoter fragment, which encodes an SD sequence near one end, and close the plasmid. This produces a set of plasmids bearing the promoter fragment separated by varying distances from the gene. This means, of course, that the DNA encoding the bacterial SD sequence will also be separated by varying distances from the ATG at the start of the gene.

How do we recognize those placements of the *lac* promoter fragment that elicit efficient expression of the desired gene? It is possible in some cases to recognize expression of the gene by functional or immunological assay of bacterial clones [see (*10*) and (*11*)]. However, we have designed a strategy that is applicable to a cloned gene even in the absence of an assay for the gene product (*12*) (Fig. 1). Our method exploits the properties of a specially constructed plasmid (pLG) bearing a portion of the *lacZ* gene of *E. coli*. This *lacZ* DNA encodes a large COOH-terminal fragment of β-galactosidase which is enzymatically active regardless of the sequence of amino acids fused to its NH_2-terminus (*13*). The plasmid does not direct expression of β-galactosidase, however, because the DNA encoding the promoter

The authors are research investigators at the Biological Laboratories, Harvard University, Cambridge, Massachusetts 02138.

0036-8075/80/0919-1428$00.75/0

and start point of translation have been deleted. As shown in Fig. 1, we first fuse a 5′ portion of the eukaryotic (or prokaryotic) gene to be expressed—called gene *X*—to plasmid pLG or a derivative thereof, so that an in-frame fusion protein is encoded (see below). If *X* is a eukaryotic gene, the fused gene *X′-′Z* is usually neither transcribed nor translated. Figure 1B shows how restriction enzymes and nucleases are used as outlined above to position the portable promoter at varying distances from the ATG encoding the NH_2-terminal methionine of protein X. A plasmid bearing an optimally positioned promoter directs synthesis of the enzymatically active X′-β-galactosidase fused protein. This protein begins with the NH_2-terminal methionine of protein X. These desired plasmids are recognized by transforming Lac⁻ bacteria with the products of the reaction of Fig. 1B and picking clones that score strongly Lac⁺ on the appropriate lactose indicator plates. Plasmids from clones producing high levels of the hybrid protein are thus recovered, and the eukaryotic gene is reconstituted. This is done by replacing the *lac*Z part of the hybrid gene with the 3′ portion of the eukaryotic gene by means of recombination in vitro (Fig. 1C). We can then monitor expression of the unfused eukaryotic protein by, for example, specific incorporation of radioactive amino acids into plasmid-encoded proteins by the "maxicell" technique (*14*) (Fig. 2). [The *lac*Z plasmid, pLG, actually bears a *lac*I-*lac*Z fusion and comes in three forms (pLG-200, -300, and -400). Each bears a unique restriction cut just 5′ to the *lac*Z coding sequence. When cut with the appropriate restriction enzyme, each plasmid is opened in a different reading frame so that if the appropriate plasmid is chosen, any coding sequence can be fused in frame with *lac*Z (see *12*).]

We have produced bacterial clones that express (separately) four eukaryotic proteins: t antigen of SV40, rabbit β-globin, and human fibroblast interferon (F-IF) with (pre-F-IF) and without (F-IF) its NH_2-terminal signal peptide (*8, 12, 15*). The sequences around the ribosome binding sites of genes that direct synthesis of the various eukaryotic proteins are shown in Fig. 3. In each case, a hybrid ribosome binding site has been formed in which the number of base pairs separating the SD sequence and the ATG (between 7 and 11 nucleotides) matches that of some known *E. coli* ribosome binding sites (*3*). In each of these cases (except t antigen), the *lac*Z fusion technique was used to recognize optimal promoter placements. In some cases, these placements were very rare. For example, in the case of F-IF, only 0.01 percent of the colonies bearing plasmids with various promoter-hybrid gene fusions were strongly Lac⁺. The fact that plasmids recognized in this way bear "hybrid" ribosomal binding sites at the beginning of the eukaryotic gene provides strong support for the hypothesis of Shine and Dalgarno concerning the key role of the SD-AUG sequences in efficient translation of mRNA.

We have estimated the amounts of protein produced by each of our plasmids in two ways. First, we have determined the levels of β-galactosidase syn-

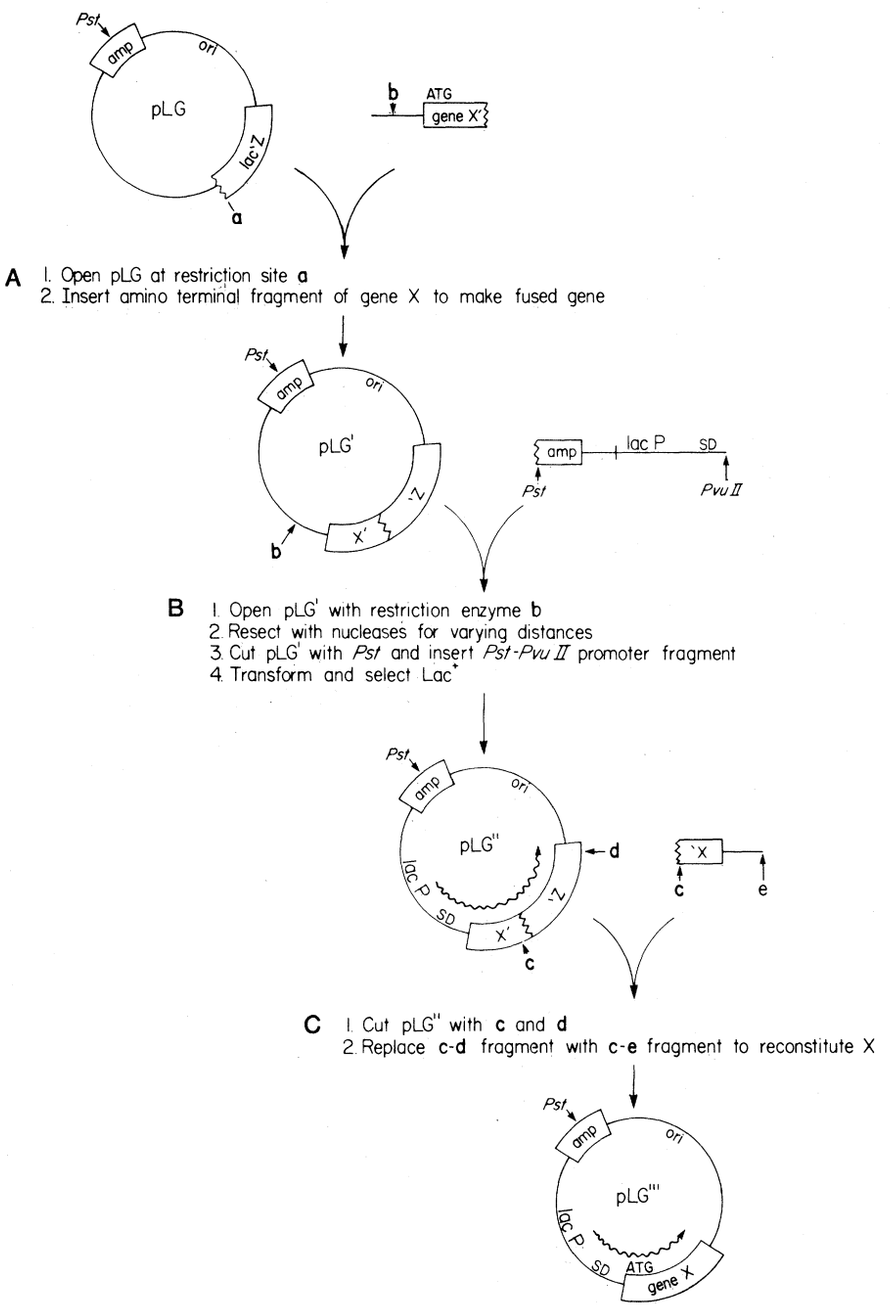
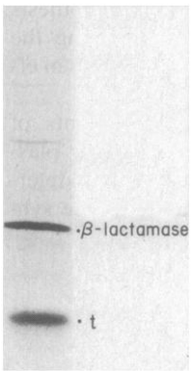


Fig. 1. A general method to maximize expression in *E. coli* of a eukaryotic gene. (A) A fragment of DNA bearing the NH_2-terminal region of gene *X* is inserted into restriction site a of a pLG plasmid, thereby fusing gene *X* to *lac*Z. (B) Plasmid pLG′ bearing the fused gene is opened at a unique restriction site b which precedes the ATG of the fused gene. Resection and insertion of a portable promoter fragment that has a Shine-Dalgarno sequence is performed. Transformed clones that bear plasmids that direct the synthesis of high levels of β-galactosidase are identified as Lac⁺ colonies on the appropriate indicator plates. (C) Gene *X* is reconstituted from plasmids that direct the synthesis of high levels of β-galactosidase as a fused product. This can be carried out, for example, by cutting the hybrid gene plasmid at any site (c) present in the gene *X* portion of the hybrid and at another site (d) in *lac*Z. Gene *X* is then reconstituted by the insertion of a DNA fragment that contains the COOH-terminal region of gene *X* extending from site c to a site (e) past the end of the gene.



.....TAACAATTTCACAC	AGGA | AACAG CT	ATG ... β-galactosidase
.....TAACAATTTCACAC	AGGA | AACAG AAAG	ATG ... SV40 t antigen (pTR436)
.....TAACAATTTCACAC	AGGA | AACAG ACAGA	ATG ... rabbit β-globin (pLG302-3)
.....TAACAATTTCACAC	AGGA | AACAG AC	ATG ... preFIF (pLG104)
.....TAACAATTTCACAC	AGGA | AACAG CC	ATG ... FIF (pLG117)

Fig. 2 (left). SV40 t antigen synthesized in bacteria. The maxicell technique (see text) was used to specifically label plasmid-encoded proteins with radioisotopes. After the cells were labeled, they were disrupted, and the contents were examined directly by polyacrylamide gel electrophoresis and autoradiography. The positions of the β-lactamase (29 kilodaltons) and SV40 small t antigen (20 kilodaltons) are indicated. Fig. 3 (right). The DNA sequences (of one strand) around the regions encoding the hybrid ribosome binding sites of several eukaryotic genes expressed in *E. coli*. The top line shows the sequence of the corresponding region of *lac*Z. The *lac*Z SD sequence is boxed as are the protein initiating ATG's. Sequences to the right of the vertical lines are from the indicated eukaryotic gene and the sequence to the left is from the portable promoter fund. The plasmids bearing the eukaryotic gene are pTR436 (*8*), pLG302-3 (*12*), pLG104 (*15*), and pLG117 (*15*).

thesis directed by the plasmid-borne eukaryotic-*lac*Z fused genes. Second, following reconstitution of the eukaryotic gene, we have compared the amount of the unfused eukaryotic proteins synthesized with that of β-lactamase in so-called maxicell experiments. [As described by Sancar *et al.* (*14*), the maxicell procedure specifically labels with radioisotopes plasmid-encoded proteins. These proteins are then readily visualized by gel electrophoresis and autoradiography.] These methods give comparable results and indicate that our strains usually produce 5,000 to 15,000 molecules of the eukaryotic protein per cell.

In two cases, the identity of the eukaryotic proteins produced in bacteria was confirmed by automated amino acid sequence analysis of radioactively labeled proteins. Both t antigen and β-globin synthesized in bacteria were found to retain their NH_2-terminal methionine. The t antigen bears an NH_2-terminal methionine when produced in animal cells, but β-globin produced in the rabbit does not.

Many proteins are normally synthesized as precursors containing NH_2-terminal signal sequences that are cleaved as the protein is excreted. Our methods will readily produce the precursor forms of these proteins. It has been reported that rat preproinsulin synthesized in bacteria was converted to the mature form by these bacteria (*16*). In the case we have examined, human pre-F-IF, we failed to detect correct processing of the precursor protein. If the precursor form of a protein is not efficiently processed by bacteria, it might be necessary to add (by DNA synthesis) an appropriately positioned ATG codon so that the mature form can be expressed directly using a bacterial promoter and a hybrid ribosome binding site (*17*). In the case of human F-IF, we were able to express the final form (F-IF) directly because the processed form of the protein begins with methionine.

The four eukaryotic proteins made by our method differ in stability in the bacterium. Pre-F-IF is quite unstable, β-globin is as stable as bacterial β-lactamase, and t antigen and mature F-IF are intermediate in stability. Extracts of bacteria producing F-IF, but not those producing pre-F-IF, display antiviral activity characteristic of authentic human F-IF (*15*). It is possible that differential sensitivity to proteases accounts for this result [see (*15*)].

There are several factors that might influence mRNA translation that we have not systematically investigated. Among these are the following: the identity of bases in the leader, including those between the SD and the AUG, and the identity of bases in the coding sequence that might effect codon usage or mRNA secondary structure. We know, for example, that in the case of the *cro* gene of phage λ, slight changes in the leader in the region 5′ to the SD had a large influence on the level of gene expression (*7*). We are now in a position to systematically analyze these effects by isolating mutations that increase expression of *lac*Z fused genes.

References and Notes

1. J. A Steitz, in *The Ribosomes: Structure, Function, and Genetics*, G. Chambliss *et al.*, Eds. (University Park Press, Baltimore, 1979), pp. 479–495; F. Sherman, J. W. Stewart, A. M. Shweingruber, *Cell* **20**, 215 (1980); M. Kozak, *ibid.* **15**, 1109 (1978).
2. J. Shine and L. Dalgarno, *Nature (London)* **254**, 34 (1975).
3. J. A. Steitz, in *Biological Regulation and Development*, R. F. Goldberger, Ed. (Plenum, New York, 1979), vol. 1, pp. 349–389.
4. K. Backman and M. Ptashne, *Cell* **13**, 65 (1978).
5. J. H. Miller and W. S. Reznikoff, Eds., *The Operon* (Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y., 1978).
6. Our original goal in this research was to produce plasmids that would direct the synthesis of large amounts of the λ repressor and cro protein. These proteins play key roles in determining the alternative modes of growth (lytic as compared to lysogenic) of the phage. We were driven to the "hybrid ribosome binding site" solution for expressing the repressor when many alternative strategies failed (*4*). The availability of strains that make large amounts of repressor and cro has greatly facilitated isolation and analysis of the properties of these proteins [see, for example (*18*)]. Moreover, the fact that production of repressor and cro may be varied over a wide range in vivo has greatly facilitated unraveling the complexities of gene control in phage λ (*19, 20*).
7. T. M. Roberts, R. Kacich, M. Ptashne, *Proc. Natl. Acad. Sci. U.S.A.* **76**, 760 (1979).
8. T. M. Roberts. I. Bikel, R. R. Yocum, D. M. Livingston, M. Ptashne, *ibid.*, p. 5596.
9. R. H. Sheller, R. E. Dickerson, H. W. Boyer, A. D. Riggs, K. Itakura, *Science* **196**, 177 (1977).
10. K. Struhl and R. W. Davis, *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5255 (1977); A. C. Y. Chang, J. H. Nunberg, R. J. Kaufman, H. A. Erlich, R. T. Schimke, S. N. Cohen, *Nature (London)* **275**, 617 (1978).
11. S. Broome and W. Gilbert, *Proc. Natl. Acad. Sci. U.S.A.* **75**, 2746 (1978).
12. L. Guarente, G. Lauer, T. Roberts, M. Ptashne, *Cell* **20**, 543 (1980).
13. P. Bassford *et al.*, in *The Operon*, J. H. Miller and W. S. Reznikoff, Eds. (Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y., 1978), pp. 245–262.
14. A. Sancar, A. Hack, D. Rupp, *J. Bacteriol.* **137**, 692 (1979).
15. T. Taniguchi, L. Guarente, T. M. Roberts, D. Kimelman, J. Douhan III, M. Ptashne, *Proc. Natl. Acad. Sci. U.S.A.*, in press.
16. K. Talmadge, J. Kaufman, W. Gilbert, *ibid.* **77**, 3988 (1980).
17. D. V. Goeddel *et al.*, *Nature (London)* **281**, 544 (1979).
18. A. D. Johnson, C. O. Pabo, R. T. Sauer, *Methods Enzymol.* **65**, 839 (1980).
19. M. Ptashne, A. Jeffrey, A. D. Johnson, R. Maurer, B. J. Meyer, C. O. Pabo, T. M. Roberts, R. T. Sauer, *Cell* **19**, 1 (1980).
20. R. Maurer, B. J. Meyer, M. Ptashne, *J. Mol. Biol.* **139**, 147 (1980); B. J. Meyer, R. Maurer, M. Ptashne, *ibid.*, p. 163; B. J. Meyer and M. Ptashne, *ibid.*, p. 195.

21. We thank A. Johnson, Carl Pabo, and R. Brent for comments on the manuscript. L.G. is supported by the Jane Coffin Childs Memorial Fund for Medical Research.

25 July 1980