

Multiple joined genes prevent product degradation in *Escherichia coli*

(multidomain/proinsulin/tandem copies/stable product)

SHI-HSIANG SHEN

Connaught Research Institute, 1755 Steeles Avenue West, Willowdale, Ontario, Canada M2R 3T4

Communicated by David R. Davies, March 6, 1984

ABSTRACT A method is described that allows the expression of a stable human proinsulin product in *Escherichia coli* as encoded by either a fused or an unfused gene construction. In the fused system, the human proinsulin coding sequence is joined to the 3' side of a fragment containing the *lac* promoter and the coding sequence for a small part of the NH₂ terminus of β -galactosidase. In the unfused system, the proinsulin coding sequence is linked directly to a fragment containing the *Tac* promoter followed by a bacterial Shine-Dalgarno sequence. In both systems, the human proinsulin product is too unstable to be detected by NaDodSO₄/polyacrylamide gel electrophoresis or even pulse-chase analysis. However, when multiple copies of the proinsulin coding sequence are tandemly linked such that the resultant protein product contains multiple copies of the proinsulin domain, the stability of the product is markedly increased in both the fused and the unfused expression systems. In the unfused system, three tandemly linked proinsulin polypeptide domains are required for stabilization, whereas two proinsulin domains plus the bacterial leader protein enhance stability in the fused system. The polypeptide product of a multiple copy proinsulin gene can be cleaved into single proinsulin units by cyanogen bromide treatment.

Recombinant DNA technology has been used extensively for the development of bacterial strains expressing useful eukaryotic products such as human insulin (1, 2), human growth hormone (3), interferon (4), and viral vaccines (5-7). However, the large-scale production of certain eukaryotic products has often been limited because of their instability in the bacterial host (4, 7, 8). It has been suggested that many eukaryotic foreign peptides are recognized as abnormal proteins in *Escherichia coli* and consequently are degraded (9, 10). For example, the half-life of human proinsulin in *E. coli* has been reported to be 2 min (11). A common strategy to circumvent this problem has been to fuse the coding sequence of the desired product to that of a host structural gene, resulting in the expression of a hybrid polypeptide. Such an expression system may provide a "native" portion added to the foreign product, thus preventing its degradation. Recently, insulin, proinsulin, and a variety of other eukaryotic proteins and viral vaccines have been produced by this method (1, 2, 12, 13). However, a major disadvantage of this approach is that the desired product constitutes only a small portion of the hybrid polypeptide, resulting in reduced yield and increased difficulties in purification. Efforts to reduce the prokaryotic moiety to a small portion of the hybrid product also usually render the product unstable (13).

A strategy is described here to prevent the degradation of human proinsulin in *E. coli* by amplifying the proinsulin coding sequence in the expression plasmid in such a way that it results in a multidomain polypeptide. It is reported that such

a multidomain polypeptide is stable in *E. coli* and can be quantitatively converted to its monomer proinsulin units by cyanogen bromide cleavage. A detailed description of the production of authentic human insulin by this system will be published elsewhere.

MATERIALS AND METHODS

Materials. Enzymes were purchased from New England Biolabs and Bethesda Research Laboratories. L-[³⁵S]Methionine (1210 Ci/mmol; 1 Ci = 37 GBq) and L-[³⁵S]cysteine (1000 Ci/mmol) were obtained from Amersham.

Bacterial Strains and Recombinant Plasmids. *E. coli* K-12 strain JM103 [$\Delta(lac\ pro)$, *thi*, *strA*, *supE*, *endA*, *sbcB*, *hsdR*, *F' traD36*, *proAB*, *lacI^q*, *Z Δ M15*] was used for all proinsulin expression experiments. Plasmid pAT/PI is a derivative of pBCA4 that carries the entire human proinsulin gene on a *Bam*HI/*Eco*RI fragment (14). The original *Eco*RI linker that precedes the initiation codon was converted from C-C-G-G-A-A-T-T-C-C-G-G to the structure A-G-A-A-T-T-C-T, to provide a modified proinsulin sequence that would maintain the desired reading frame in subsequent steps. The resulting modified proinsulin sequence on a *Bam*HI/*Eco*RI fragment was inserted into plasmid pAT153 between the *Eco*RI and *Bam*HI sites (unpublished data). Plasmid pTac bears a *Tac* promoter sequence modified from plasmid pDR540 (15) in the following way: pDR540 DNA was cut with *Bam*HI and *Hind*III and digested with mung bean nuclease to remove the sticky ends. The isolated *Tac* promoter fragment was inserted into the vector plac 504/PI, which had been digested with *Eco*RI and *Hind*III and filled in with DNA polymerase I (Klenow fragment). The resulting construct was designated as pTac. Details concerning the construction of plac 504/PI will be published elsewhere. Plasmid plac 239, to be described in detail elsewhere, carries a *lac* promoter and part of the *lacZ* gene encoding 80 amino acids of the NH₂ terminus of β -galactosidase.

Purification of the Proinsulin Product and Polyacrylamide Gel Electrophoresis. Bacterial cultures (3.0 ml) were grown in YT medium and induced at an approximate cell density of 0.1 OD₅₆₀ unit by the addition of isopropyl β -D-thiogalactoside to 1 mM. After 12 hr, the cells were harvested by centrifugation. Cell pellets were either dissolved directly in the sample buffer and analyzed by NaDodSO₄/polyacrylamide gel electrophoresis as described by Laemmli (16) or further purified as follows: cells were suspended in 1.0 ml of the sonication buffer as described by Bikel *et al.* (17) and subjected to sonic disruption. The sonicated material was centrifuged. The pellets were analyzed by NaDodSO₄/polyacrylamide gel electrophoresis as described above.

Pulse-Chase Experiments. Bacterial cultures (5.0 ml) were grown in minimal salts medium/0.2% glucose/0.001% thiamin at 37°C. At a cell density of 0.5 OD₅₆₀ unit, the cultures were induced as before and, after 5 min, 300 μ Ci each of L-[³⁵S]methionine and L-[³⁵S]cysteine were added. Labeling

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.

proceeded for 30 sec, after which incorporation of radioactivity was chased by the addition of 100 μ l of 500 mM unlabeled L-methionine/L-cysteine. Aliquots (0.5 ml) of each culture were removed at predetermined times and immediately centrifuged for 10 sec, and cells were sonicated as described above.

Cyanogen Bromide Cleavage. The proinsulin polypeptide purified by sonication was dissolved in 20 ml of 70% formic acid and cleaved by treatment with 50 mg of cyanogen bro-

midide per mg of protein for 35 hr at room temperature. Products were analyzed by NaDodSO₄/polyacrylamide gel electrophoresis.

RESULTS

Construction of the Multiple Proinsulin Genes. The strategy for the construction of the multidomain proinsulin genes in both the fused and unfused systems is diagrammed in Fig. 1.

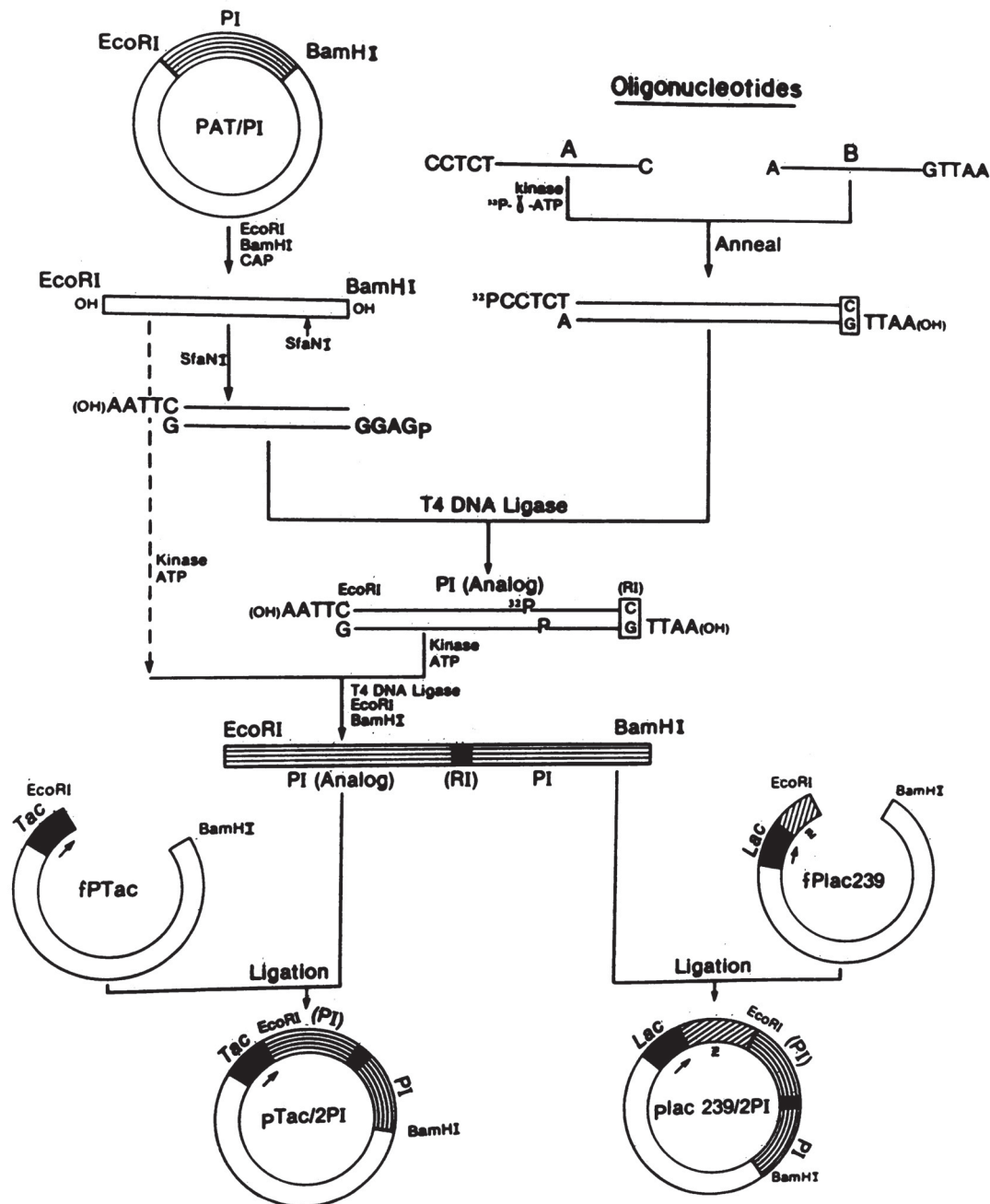
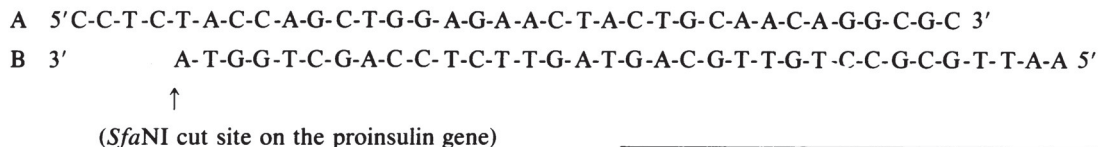
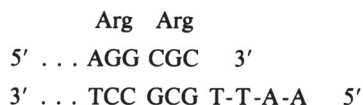


FIG. 1. Construction of the multiple proinsulin genes. The multiple proinsulin gene was isolated from an agarose gel and inserted into the vectors fPTac and fplac 239. fPTac is an EcoRI/BamHI fragment of pTac. The construction was arranged to give an EcoRI site immediately 3' to the promoter. fplac 239 is an EcoRI/BamHI fragment of plac 239. The construction was first arranged to give an EcoRI site at amino acid 80 in the lacZ gene for insertion of multiple genes and then to restore the proper reading frame of the proinsulin coding sequences by cleaving at the EcoRI site of the lac-proinsulin junction, digesting with mung bean nuclease, and recircularizing by blunt-end ligation.

A and B are synthetic oligonucleotides having the following sequences:



Oligonucleotide A was first phosphorylated and then hybridized to oligonucleotide B. After hybridization, the two oligonucleotides form a part of the human proinsulin coding sequence (14) phosphorylated on the end beginning at the *Sfa*NI cut site near the 3' end of the coding sequence and terminating with the last proinsulin codon AAC (asparagine) followed by the unphosphorylated additional sequence



The extra sequence of the synthetic oligonucleotides can encode two additional arginine residues and also provides a sticky end for ligation to an *Eco*RI sticky end. Note that such a ligation destroys the *Eco*RI site because of the C-G base pair (boxed in Fig. 1).

The synthetic human proinsulin gene was removed from pAT/PI by *Eco*RI/*Bam*HI digestion, dephosphorylated, and further digested with *Sfa*NI. The asymmetrically phosphorylated proinsulin gene fragment enables unidirectional ligation to the A and B oligonucleotides, resulting in the proinsulin gene analog [PI (analog) in Fig. 1]. The proinsulin gene analog was joined to another proinsulin gene sequence (*Eco*RI/*Bam*HI fragment) in a controlled unidirectional ligation reaction carried out with T4 DNA ligase in the presence of *Eco*RI and *Bam*HI. The resulting double proinsulin gene fragment was further ligated to additional proinsulin gene analogs to give plasmids containing up to seven joined proinsulin coding sequences.

Joined proinsulin gene fragments containing different numbers of the proinsulin coding sequence were inserted into one or the other of two vectors, pTac and plac 239 (to be described in detail elsewhere). In the former case, a *Tac* promoter followed by a bacterial Shine-Dalgarno sequence is joined directly to the proinsulin coding sequence with its own initiator codon. The resulting plasmids give products containing one or more proinsulin domains. pTac/2PI in Fig. 1 has two such domains. In the latter case, the proinsulin coding sequence is preceded by the *lac* promoter and 80 codons from the 5' end of the β -galactosidase gene. The resulting plasmids give fused gene products (part of β -galactosidase fused to one or more domains of proinsulin). plac 239/2PI in Fig. 1 has two proinsulin domains fused to 80 amino acids of β -galactosidase. The various constructs, containing one to seven copies of the proinsulin coding sequence, were confirmed by restriction mapping and DNA sequence analysis.

Expression of Multiple Proinsulin Genes in *E. coli*. The multidomain proinsulin gene constructs, fused to β -galactosidase and unfused, were expressed in *E. coli* strain JM103 after induction, and the resultant products were analyzed by NaDodSO₄/polyacrylamide gel electrophoresis as shown in Fig. 2. In the unfused system (Fig. 2A), the product of one proinsulin coding sequence is too little to be visualized in a stained gel after NaDodSO₄/polyacrylamide gel analysis (lane 1). However, the level of product is strikingly enhanced as the proinsulin coding sequence copy number is

increased. The product of two joined coding sequences is visible (lane 2) after removal of some of the bacterial pro-

teins by sonication, while the amount of product with three joined coding sequences is considerably increased (lane 3). A further increase in the number of joined coding sequences up to five does not appear to cause a significantly greater quantity of product (lane 5). A definite decrease is seen with seven joined coding sequences (data not shown).

Similar results were obtained in the fused expression system in which the first proinsulin coding sequence was preceded by 80 codons of the *lacZ* gene (Fig. 2B). The fusion product with a single proinsulin coding sequence was difficult to visualize (lane 1) although, after partial purification by sonication, a light band could be detected (lane 4) by loading twice as much sample as in lane 1. Similarly to the unfused system, the amount of product obtained in the fused system was greatly enhanced by increasing the number of joined proinsulin domains to two or more (lanes 2 and 3 as well as 5 and 6).

Stability of Products. As described above, the appearance of prominent bands on NaDodSO₄/polyacrylamide gel electrophoresis indicates that the multidomain proinsulin products are present in markedly greater amounts than their monomeric counterparts. A possible cause of this result is a greater stability of the multidomain polypeptide. To investigate this factor, pulse-chase experiments were carried out. As shown in Fig. 3, the expected product from bacteria harboring a single copy of the proinsulin coding sequence (plasmid pTac/PI) could not be detected (lanes 1-4), even with a chase as short as 30 sec. The product from two joined proin-

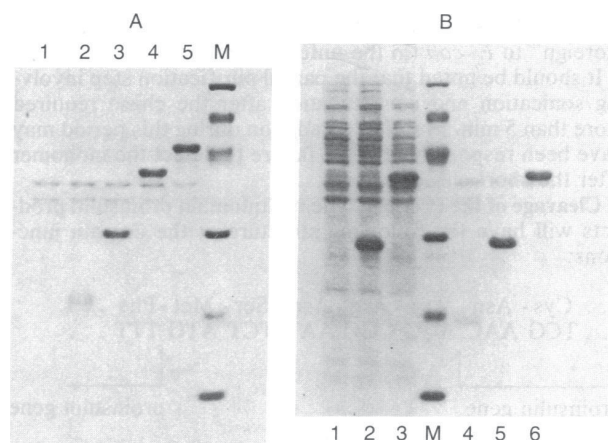


FIG. 2. Electrophoretic analysis of multiply expressed proinsulin polypeptides. Proinsulin polypeptide preparations were analyzed on NaDodSO₄/15% polyacrylamide gels. (A) Partially purified products of the unfused system construction. Lanes 1-5 represent products equivalent to 500 μ l of original culture from plasmids pTac/PI, pTac/2PI, pTac/3PI, pTac/4PI, and pTac/5PI, respectively. (B) Products of the fused construction system. Lanes 1-3 represent total cell protein equivalent to 150 μ l of original culture bacteria containing plasmids plac 239/PI, plac 239/2PI, and plac 239/3PI, respectively. Lanes 4-6 represent products from the same plasmids as in lanes 1-3 after partial purification by sonication. Lanes M: protein molecular weight markers: from top to bottom, 94,000, 67,000, 30,000, 20,100 and 14,400.

ing sequences were both found to be the same—approximately 120 min. Why increasing the number of joined sequences to seven resulted in a lower yield remains an interesting problem for study. Also, it would be interesting to determine whether changes in the number of joined proinsulin coding sequences alter transcriptional efficiencies or mRNA stability.

It has been reported that disulfide bond formation does not occur in the cytoplasm of *E. coli* because of a low electrochemical potential in the cell (18). Thus, the product stability observed with tandemly repeated proinsulin domains is unlikely due to increased numbers of intramolecular disulfide bridges. I have observed, by electron microscopy, inclusion-like bodies in cells producing stable tandemly linked proinsulin polypeptide but could not find such structure in cells producing an unstable single-domain polypeptide product. Such inclusion-like bodies may result from the aggregation of polypeptide products of limited solubility within the cell. A similar interpretation was proposed concerning the formation of inclusion-like bodies by such aggregation in production of the recombinant insulin polypeptide of *E. coli* (19). Consequently, I suggest that the aggregation of the tandemly linked proinsulin polypeptide products sequestered into the insoluble inclusion-like bodies could avoid an attack from proteolytic enzymes in *E. coli*. A similar suggestion of product aggregation has been made for the stabilization of a degradable protein-X90 (20). The formation of protein aggregates may require not only a certain protein concentration but also a critical length of individual polypeptide units. Achieving such conditions could be accomplished by either joining single polypeptide coding sequences together or fusing a coding sequence to a bacterial leader of sufficient length as discussed for the construction. The stabilization of somatostatin expression in *E. coli* was achieved by the latter strategy (21).

The mechanisms by which abnormal proteins, including foreign polypeptides, are degraded in *E. coli* are not fully understood. The *E. coli* lon protease, a DNA-binding protein with ATP-dependent proteolytic activity, clearly plays a role in this process (22, 23). There are some *E. coli* mutants (22, 24) that stabilize some otherwise unstable polypeptides and nonsense fragments. However, these mutants do not stabilize a single proinsulin polypeptide in *E. coli* (25).

The method described here for stabilizing an expressed polypeptide via a multidomain joined product could be applicable to the expression of other unstable products. It is anticipated that both the range of desirable peptides and the availability of methods for sequence-specific cleavage of peptide bonds will increase, thus improving the general applicability of the multidomain polymer strategy here described.

I am grateful to Dr. Keith Dorrington for his encouragement and critical reading of the manuscript, to Dr. Eric James for valuable discussions, and to Dr. Oliver Smithies for his valuable comments

on this work. I thank Dr. Joel Haynes for oligonucleotide synthesis and comments on the manuscript, Dr. Robert Garvin for his comments on this work, and Dr. K. Tsai for electron microscopy observation. I gratefully acknowledge Richard Elliott, Cheryl Davey, and Elizabeth Ciemniowski for their assistance and Lynne-Marie McKay for her efforts in preparing the manuscript. This work was supported by a Programme for Industry/Laboratory Projects contract from the National Research Council of Canada.

1. Chan, S. J., Weiss, J., Konrad, M., White, T., Bahl, C., Yu, S.-D., Marks, D. & Steiner, D. F. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 5401–5405.
2. Goeddel, D. V., Kleid, D. G., Bolivar, F., Heyneker, H. L., Yansura, D. G., Crea, R., Hirose, T., Kraszewski, A., Itakura, K. & Riggs, A. D. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 106–110.
3. Goeddel, D. V., Heyneker, H. L., Hozumi, T., Arentzen, R., Itakura, K., Yansura, D. G., Ross, M. J., Miozzari, G., Crea, R. & Seeburg, P. H. (1979) *Nature (London)* **281**, 544–548.
4. Taniguchi, T., Guarente, L., Roberts, T. M., Kimelman, D., Douhan, J., III, & Ptashne, M. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 5230–5233.
5. Yelverton, E., Norton, S., Obijeski, J. F. & Goeddel, D. V. (1983) *Science* **219**, 614–620.
6. Malek, L. T., Soostmeyer, G., Garvin, R. T. & James, E. (1984) in *Modern Approaches to Vaccines*, eds. Chanock, R. M. & Lerner, R. A. (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY), pp. 203–208.
7. Davis, A. R., Bos, T., Ueda, M., Nayak, D. P., Dowbenko, D. & Compans, R. W. (1983) *Gene* **21**, 273–284.
8. Itakura, K., Hirose, T., Crea, R., Riggs, A. D., Heyneker, H. L., Bolivar, F. & Boyer, H. W. (1977) *Science* **198**, 1056–1063.
9. Goldschmidt, R. (1970) *Nature (London)* **228**, 1151–1154.
10. Lin, S. & Zabin, I. (1972) *J. Biol. Chem.* **247**, 2205–2211.
11. Talmadge, K. & Gilbert, W. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 1830–1833.
12. Shine, J., Fettes, I., Lan, N. C. Y., Roberts, J. L. & Baxter, J. D. (1980) *Nature (London)* **285**, 456–461.
13. Reed, S. I. (1982) *Gene* **20**, 255–265.
14. Brosseau, R., Scarpulla, R., Sung, W., Hsiung, H. M., Narang, S. A. & Wu, R. (1982) *Gene* **17**, 279–289.
15. Russell, D. R. & Bennett, G. N. (1982) *Gene* **20**, 231–243.
16. Laemmli, U. K. (1970) *Nature (London)* **227**, 680–685.
17. Bikel, I., Roberts, T. M., Bladon, M. T., Green, R., Amann, E. & Livingston, D. M. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 906–910.
18. Pollitt, S. & Zalkin, H. (1983) *J. Bacteriol.* **153**, 27–32.
19. Williams, D. C., Van Frank, R. M., Muth, W. L. & Burnett, J. P. (1982) *Science* **215**, 687–689.
20. Cheng, Y.-S. E., Kwok, D. Y., Kwok, T. J., Soltvedt, B. C. & Zipser, D. (1981) *Gene* **14**, 121–130.
21. Itakura, K., Hirose, T., Crea, R. & Riggs, A. D. (1977) *Science* **198**, 1056–1063.
22. Gottesman, S. & Zipser, D. (1978) *J. Bacteriol.* **133**, 844–851.
23. Charette, M. F., Henderson, G. W. & Markovitz, A. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 4728–4732.
24. Bukhari, A. I. & Zipser, D. (1973) *Nature (London) New Biol.* **243**, 238–241.
25. Emerick, A. W., Bertolani, B. L., Ben-Bassat, A., White, T. J. & Konrad, M. W. (1984) *Biotechnology* **2**, 165–168.