

Primary structure of the gene encoding rat preprosomatostatin

(somatostatin/gene structure/promoter)

MARC R. MONTMINY*[†], RICHARD H. GOODMAN*[†], SHARON J. HOROVITCH[†], AND JOEL F. HABENER*

*Laboratory of Molecular Endocrinology, Massachusetts General Hospital, and Howard Hughes Medical Institute Laboratories, Harvard Medical School, Boston, MA 02114; and [†]Division of Endocrinology, Department of Medicine, Tufts-New England Medical Center, Boston, MA 02111

Communicated by Louis B. Flexner, February 14, 1984

ABSTRACT The somatostatins are peptides of 14 and 28 amino acids that are produced in a variety of endocrine and nonendocrine tissues. These peptides inhibit the secretion of many different pituitary, pancreatic, and gastrointestinal hormones. Previously, we have reported the isolation and nucleotide sequence of a cDNA derived from a rat medullary thyroid carcinoma that encoded preprosomatostatin, a 116-amino-acid precursor of somatostatin. We now report the structural characterization of the rat somatostatin gene isolated from recombinant bacteriophage libraries prepared from rat liver DNA. The gene spans 1.2 kilobases and is interrupted within the coding sequence of prosomatostatin by a single intron of 630 bases. A sequence characteristic of a Goldberg-Hogness promoter ("TATA" box), T-T-T-A-A-A, is located 31 bases upstream from the transcriptional initiation site. A repetitive DNA sequence, highly reiterated in the rat genome, is located in the 5' flanking region of the gene within 900 bases of the initiation site.

The somatostatins are peptides of 14 and 28 amino acids that are produced in a variety of endocrine and nonendocrine tissues, including the hypothalamus, cerebral cortex, pancreas, stomach, and duodenum (1, 2). Both somatostatin-28 and -14 are biologically active, but tissues vary in their ability to generate one or the other peptide (3, 4). The major actions of somatostatin-28 and -14 are the inhibition of secretion of a variety of peptide hormones, including growth hormone, thyrotropin, glucagon, and insulin (5). The rat somatostatin gene encodes preprosomatostatin, a precursor of 116 amino acids that is processed cotranslationally within the endoplasmic reticulum to yield prosomatostatin, a peptide of 92 amino acids (6). Prosomatostatin is subsequently cleaved post-translationally to produce somatostatin-28 and somatostatin-14, peptides that correspond to the carboxyl-terminal 28 and 14 amino acids of prosomatostatin. Other peptides generated from prosomatostatin include a peptide corresponding to the amino-terminal 12 amino acids of somatostatin-28 and peptides containing sequences within the amino-terminal extension of prosomatostatin (7, 8). The biological function of these peptides is unknown, but the high degree of conservation of their amino acid sequences within different species (6, 9, 10) and their transport to and localization within nerve endings (7) suggest that they may also serve some biological function.

To facilitate the identification of factors that regulate somatostatin gene expression and post-translational processing, it is useful to elucidate the structure of the somatostatin gene. Towards this end, we have isolated and structurally characterized cloned genomic sequences that contain the rat somatostatin gene.

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.

METHODS AND MATERIALS

Screening Genomic Libraries. Somatostatin cDNA was excised from the plasmid 1-6-13 (6) by digestion with *Pst* I. The excised cDNA insert fragments were labeled with [³²P]dCTP by nick-translation (11) to a specific activity of 1–3 × 10⁸ cpm/μg of DNA. Partial *Eco*RI and partial *Hae* III Charon 4A bacteriophage libraries obtained from T. Sargent, R. Wallace, and J. Bonner (California Institute of Technology) (12) were screened by the method of Benton and Davis (13).

Individual hybridizing plaques were purified, the bacteriophage clones were grown, and phage DNA was isolated as described by Blattner *et al.* (14).

Subcloning. DNA fragments to be subcloned were ligated to pBR322 by using T4 DNA ligase (Boehringer). The ligated products were used to transform *Escherichia coli* strain C600. Recombinant colonies were grown on agar plates supplemented with the appropriate antibiotics and screened with ³²P-labeled DNA fragments (15). Hybridizing colonies were picked and grown as described (6).

Southern Blot Analysis. Genomic DNA prepared from a single rat spleen (16) was digested with various restriction endonucleases. The DNA digests were electrophoresed on 0.8% agarose gels at 30 V overnight, transferred to nitrocellulose filters, and hybridized with ³²P-labeled nick-translated cDNA probes according to the method of Southern (17).

Mapping mRNA 5' Termini. RNA was isolated from rat medullary thyroid carcinoma tissue as described by Chirgwin *et al.* (18). RNA preparations were enriched for the polyadenylated fraction by passage through an oligo(dT)-cellulose column. S1 nuclease mapping (19) was performed with the subclone RSTs-1 (Fig. 1). The plasmid was digested with the restriction enzyme *Bss*H2 and exchange-labeled by using polynucleotide kinase. The labeled probe was denatured and added to the RNA, and the mixture was allowed to hybridize overnight in 80% formamide at 57°C. Digestions were performed with 55 units of S1 nuclease (Boehringer) for 40 min at 30°C. The digestion products were analyzed on an 8% denaturing polyacrylamide gel along with a sequencing ladder to determine the length of the S1 nuclease-protected fragment.

Sequence Determination. Nucleotide sequence was determined according to the method of Maxam and Gilbert (20). Restriction fragments were labeled at their 5' termini by using T4 polynucleotide kinase (Bethesda Research Laboratories) and at their 3' termini with the large fragment (Klenow) of DNA polymerase I (Boehringer).

Mapping Repetitive DNA Sequences. Three methods were used to identify middle repetitive sequences flanking the rat somatostatin gene. Initially, total rat genomic DNA, sheared by sonication to an average size of 1 kilobase, was nick-translated and hybridized to Southern blots of the rat somatostatin genomic clone cut with various restriction enzymes. Somatostatin gene fragments, which were identified as containing repetitive sequences by the above procedures, were nick-translated and hybridized to Southern blots of total genomic DNA cut with various restriction enzymes. Finally,

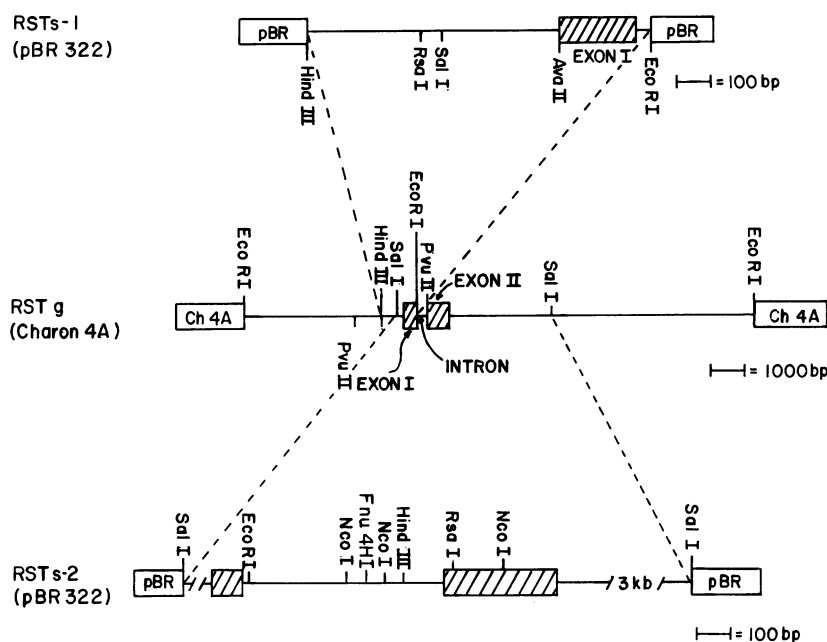


FIG. 1. Recombinant λ phages containing the rat preprosomatostatin gene (RST g) and two subclones of the genomic insert (RSTs-1 and RSTs-2) in plasmid pBR322. Recombinant bacteriophages were isolated by plaque-hybridization from cloned rat liver DNA libraries by using a recombinant plasmid encoding rat preprosomatostatin as a hybridization probe. Nucleotide sequences of the DNA inserts were determined entirely on both strands. Orientation of the rat preprosomatostatin gene is from left to right. Open boxes represent DNA of the cloning vehicles, λ phage Charon 4A and plasmid pBR322. Hatched boxes indicate the two exons of the gene. Lines represent the remaining rat DNA, including the single intron. kb, Kilobases; bp, base pairs.

these nick-translated fragments were hybridized to recombinant phage plaques containing rat genomic fragments (13). In this way, a rough estimate of the frequency of this repetitive element was obtained by determining the proportion of hybridizing bacteriophage plaques.

RESULTS

Isolation and Sequence Determination of the Rat Somatostatin Gene. The rat preprosomatostatin cDNA (6) was used as a labeled hybridization probe to screen two genomic libraries cloned in Charon 4A bacteriophages. The first library, prepared from a partial *EcoRI* digest of rat liver genomic DNA (12), contained a recombinant bacteriophage that represented the 5' exon of the rat somatostatin gene and extended 15 kilobases in the 5' direction. A second recombinant bacteriophage, isolated from the partial *Hae III* rat genomic library (12), contained preprosomatostatin gene sequences that overlapped with those in the first bacteriophage and extended 10 kilobases in the 3' direction. The restriction map of this 25-kilobase region indicated that the somatostatin gene is encompassed by two *EcoRI* fragments of 4.6 kilobases and 9.7 kilobases, which represent the 5' and 3' regions of the gene, respectively (Fig. 1). Analysis by Southern blotting of rat spleen genomic DNA revealed hybridizing fragments identical to those obtained from digests of the two recombinant bacteriophages (Fig. 2 and unpublished data). This result indicates that the bacteriophage libraries used contain an accurate representation of the somatostatin gene. Genomic blots probed with the 5' and 3' fragments of the somatostatin cDNA revealed two hybridizing bands, most consistent with the existence of a single somatostatin gene in the rat. Additional bands were seen on long exposures of the blots, however, raising the possibility of multiple somatostatin-related genomic sequences.

The two hybridizing *EcoRI* fragments isolated from the bacteriophage clones were subcloned into the plasmid pBR322 (Fig. 1). Restriction fragments were labeled either at the 3' ends with the large fragment of DNA polymerase I or at the 5' ends with polynucleotide kinase and were se-

quenced by the method of Maxam and Gilbert (20). Comparison of the nucleotide sequences of the subclones with that of the cDNA encoding rat preprosomatostatin revealed a single intron of 630 bases. This intervening sequence interrupts the coding region of preprosomatostatin within the amino-terminal peptide extension of somatostatin-28 (Fig. 3). Characteristic G-T/A-G donor-acceptor sites are present at the ends of the intervening sequence. The sequence of the preprosomatostatin coding region in the genomic subclones is identical to that of the cDNAs that have been sequenced (6, 9). The genomic clone and the cDNA were found to differ by two bases within the 3' untranslated region.

Determination of the Transcriptional Initiation Site. The point of initiation of transcription was localized by S1 nucle-

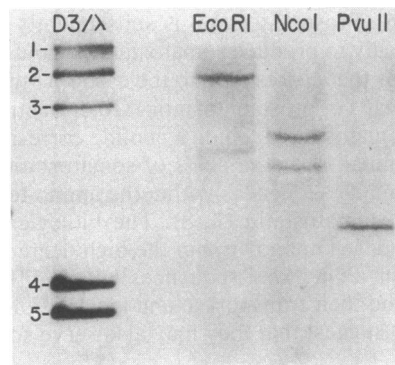


FIG. 2. Autoradiogram of genomic (Southern) blot of restriction enzyme digests of rat spleen DNA. Aliquots (10 μ g) of rat spleen DNA were digested with one of several restriction enzymes, subjected to electrophoresis through an 0.8% agarose gel, and transferred onto nitrocellulose. The blot was hybridized to a 32 P-labeled cDNA fragment encoding rat preprosomatostatin. The sizes of the hybridizing bands were estimated by a comparison with a marker lane of *HindIII*-digested λ phage DNA (D3/ λ). The respective sizes of the λ phage fragments in kilobase pairs (labeled 1-5) are 23, 9.4, 6.7, 2.3, and 2.0. An 0.6-kilobase *Pvu II* hybridizing band is not shown. Exposure time was 48 hr.

```

gaagtgaccagccgaatagccttaagcacccttgccataccacagaccgttaagcatgatggcaagtcagtaatc
tgagtacattgacaggtacccaactgtgtgtgctgatgtatgctggccaaggactgaagatctcagtaattaatcag
ccctatgtggcggaaataggatagcctgacactgagtgaaagcaagattattggctgtgtggcgtggagaa
ttcatgtgctgtgtgggtgaggcttcttttctcaaaaaaaaaaaaaataaacctttagatcgtgtgacctccc
ctcacttcttgattgattttgagaggcctaaatggtgctgtaaaagcactggtgagatctgggggcctccttgctgac
(-31) (+1)
tcagagagagagtttaaaaggggagaccgtggagcctcgaatAGCGGCTGAAGGAGACGCTACTGGAGTCGCTCTGC
-24
TCGCTGCGGACCTGCGTCTAGACTGACCCACCGGCTCAAGCTCGGCTGCTGAGGCAGGGGAGATG CTG TCC TGC
met leu ser cys

arg leu gln cys ala leu ala ala leu cys ile val leu ala leu gly gly val thr gly
CGT CTC CAG TGC GCG CTG GCC GCG CTC TGC ATC GTC CTG GCT TTG GGC GGT GTC ACC GGG
+1 +10 +20
ala pro ser asp pro arg leu arg gln phe leu gln lys ser leu ala ala ala tyr gly
GCG CCC TCG GAC CCC AGA CTC CGT CAG TTT CTG CAG AAG TCT CTG GCG GCT GCC ACC GGG

lys gln
AAA CAG gtaaggaatggctgggactcgtccctttgcaattccccggccttccccttagcttctgctgtagccctg
cgacaggtgttttagcgggcgttctcagagtcgctcagccctgagctcccagggaaactttgaagctagggtcgc
tcttactcgttccagaattgatcggcgtggtggtcaccctgcaggttaagttcccccttcgcttcaggaaaaatcccga
agcctgcaagagagcggggagagactgagctctatccctggctactggcagcaggggttctgacccaggtgctgaaaaaa
atccggcaagaactcaggtccatggtccattctgtgctcataaaggaaaatggagctctcaaaactattggcatactat
atttcaaaaacgacttctcatcattcctggtttctgtggttttaaggcatagcacttctgaaagacttgggtttgagg
aagctttttccctgtgataatcagtaataagcagccatccatattactgtgaaacttggtttgaatgattaa

glu leu ala
tcttattttcaaacccatttctccctttctccattcccccttttgctctctccctgcccattccagGAA CTG GCC
+30 +40
lys tyr phe leu ala glu leu leu ser glu pro asn gln thr glu asn asp ala leu glu
AAG TAC TTC TTG GCA GAA CTG CTG TCT GAG CCC AAC CAG ACA GAG AAC GAT GCC CTG GAG
+50 +60
pro glu asp leu pro gln ala ala glu gln asp glu met arg leu glu leu gln arg ser
CCT GAG GAT TTG CCC CAG GCA GCT GAG CAG GAC GAG ATG AGG CTG GAG CTG CAG AGG TCT
+70 +80
ala asn ser asn pro ala met ala pro arg glu arg lys ala gly cys lys asn phe phe
GCC AAC TCG AAC CCA GCC ATG GCA CCC CGG GAA CGC AAA GCT GGC TGC AAG AAC TTC TTC
+90
trp lys thr phe thr ser cys stop
TGG AAG ACA TTC ACA TCC TGT TAG CTTAATATTGTTGTTCTCAGCCAGACCTCTGATCCCCTCTCTGCAAA
TCCCATATCTCTTCTTAACTCCAGCCCCCCCCCAATGCTCAACTAGACCCTGCGTTAGAAATTGAAGACTGTAAT
TACAAAATAAAATTATGGTGAATTATGAA

```

Fig. 3. Nucleotide sequence of the rat preprosomatostatin gene. The nucleotide sequence was determined from the subcloned DNAs shown in Fig. 1. Exons of the gene are shown in capital letters, and the intron and flanking DNA are shown in lower-case letters. The promoter (Goldberg-Hogness) sequence and the A-A-T-A-A polyadenylation signal are underlined. The amino acid sequences of somatostatin-28 (SS-28) and somatostatin-14 (SS-14) are overlined and boxed, respectively. The protein-coding sequence of the gene is keyed by numbers above the assigned amino acids. Amino acid +1 (alanine) begins the sequence of prosomatostatin. Amino acid -24 (methionine) designates the beginning of the signal (leader) peptide sequence, and -31 designates the position of the promoter relative to the position of the cap site (+1).

ase mapping (19). Polyadenylated RNA from a rat medullary thyroid carcinoma that produced somatostatin (21) was hybridized to a fragment of the recombinant bacteriophage representing the 5' region of the gene. Subsequently, the hybrids were digested with S1 nuclease (19). The length of the nuclease-resistant fragment was determined by electrophoresis of the digest on an 8% sequencing gel (Fig. 4). A single band was observed that corresponded to a DNA fragment of 184 bases in length. This finding placed the site of the initiation of transcription at 102 bases upstream from the initiator methionine codon. Thirty-one bases upstream from this initiation site lies an A+T-rich sequence, T-T-T-A-A-A-A, surrounded by a G+C-rich sequence. This A+T-rich sequence is likely to represent the Goldberg-Hogness promoter ("TATA" box).

Repetitive DNA Sequences. DNA isolated from the bacteriophage clones was digested with *EcoRI*, transferred to nitrocellulose, and probed with nick-translated rat genomic DNA. These blots revealed a 4.7-kb hybridizing fragment identical to that which hybridized to the 5' portion of the somatostatin cDNA probe. A similar experiment carried out with a *HindIII-EcoRI* fragment of one of the subclones localized the repetitive element to within 900 base pairs of the transcriptional initiation point. Regions downstream from the *Sal I* site (located at position -350) did not hybridize to the nick-translated genomic DNA probe, indicating that the repetitive sequence lies between -350 and -900. To esti-

mate the frequency of this repetitive element, the *HindIII-EcoRI* fragment labeled with ³²P was hybridized to nitrocellulose filters containing phage DNA plaques that had been transferred from the rat *Hae III* library. Approximately 10%

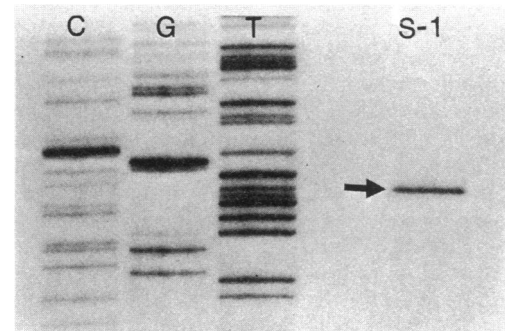


Fig. 4. Localization of the transcriptional initiation site of the rat preprosomatostatin gene by S1 nuclease mapping of a DNA-RNA hybrid. mRNA prepared from a rat medullary thyroid carcinoma was hybridized to a 5' end-labeled *BssH2* fragment of subclone RSTs-1. The hybrid was digested with S1 nuclease, and the products were analyzed on an 8% polyacrylamide/urea gel (lane labeled S-1) along with cytosine (C), guanine (G), and thymidine (T) sequencing of ladders prepared from a cDNA fragment encoding the α subunit of rat lutropin.

of the plaques hybridized strongly to the *Hind*III-*Eco*RI fragment, placing a lower limit of 20,000 copies of the repetitive element per genome. When restriction digests of genomic DNA were analyzed by Southern blot hybridization with the *Hind*III-*Eco*RI fragment as a probe, a continuous smear of hybridization was observed. No repetitive elements were detected within the intervening sequence of the rat somatostatin gene.

DISCUSSION

Sequence determinations of bacteriophage Charon 4A clones encoding rat somatostatin revealed a gene of approximately 1.2 kilobases. Only the 5' portion of the somatostatin gene was represented in the partial *Eco*RI library (12). This fragment encoded the 5' flanking region of the gene, the 5' untranslated region, the amino-terminal portion of preprosomatostatin, and a small portion of the intervening sequence (Fig. 1). No sequences representing the 3' portion of the gene were detected in this library, despite the 10 genome-equivalents that were screened. The complete rat somatostatin gene, however, was represented in the partial *Hae* III library.

Characteristic features of the rat somatostatin gene included a variant TATA box or Goldberg-Hogness promoter, T-T-T-A-A-A-A, 31 bases upstream from the transcriptional initiation site. The unique band seen in the S1 nuclease experiment suggests that transcription of the gene occurs at a single site. The 5' untranslated region of the mRNA includes an open reading frame of 34 codons which are in phase with the coding region of preprosomatostatin. No AUG codons are present within the 5' untranslated region, however.

A single intervening sequence 630 bases in length was identified, separating the Gln-Glu codons at amino acid positions 46-47. The biological function of this region of preprosomatostatin is not known. Therefore, it is not possible to assess whether the intervening sequence actually separates functional domains within the precursor or merely separates the signal peptide from the somatostatin peptides. Preliminary studies with an antibody directed towards the middle of the prosomatostatin molecule (7) suggest that prosomatostatin may be cleaved at a site amino-terminal to the somatostatin-28 cleavage site (22). Benoit *et al.*, using an antiserum to somatostatin-14, also have found evidence suggestive of cleavage within the amino-terminal region of prosomatostatin (23). The precise location of these additional cleavage sites, and consequently their relationships to the intervening sequence within preprosomatostatin, are unknown. The known cleavage sites, the Gly-Ala sequence separating the leader sequence from prosomatostatin (6), the Arg-Ser sequence separating the amino-terminal portion of prosomatostatin from somatostatin-28, and the Arg-Lys sequence separating somatostatin-14 from the amino-terminal portion of somatostatin-28 clearly are not sites that correspond to intervening sequences in the gene.

Southern blotting analysis was most consistent with the existence of a single gene encoding somatostatin in the rat. This observation is in contrast to the situation in lower species, such as fish, where at least two distinct somatostatin genes are present (24). Noe and Spiess have presented evidence that the two anglerfish preprosomatostatins are processed differently to yield somatostatin-14 and a peptide similar in size to somatostatin-28 (25). If indeed there is only a single somatostatin gene in the rat, then the ability of various tissues to produce somatostatin-14 or -28 must depend on the ability of particular cell types to process the single precursor in different ways. Similar conclusions have been reached regarding the processing of proopiomelanocortin (26).

The significance of the repetitive sequence located near the promoter region of the somatostatin gene is not known. Southern blot hybridizations indicated that this repetitive sequence lies between bases -350 to -900. Therefore, it may be too far upstream to influence somatostatin biosynthesis. Nonetheless, determination of the influence of this region may provide a first step in elucidating the factors that regulate somatostatin biosynthesis.

Note Added in Proof. Since this manuscript was submitted, we have determined the nucleotide sequence of the repetitive DNA element. The repetitive element contains a 42-base-pair region of alternating G-T sequence characteristic of DNA with Z-forming potential.

We thank Esther G. Hoomis for excellent secretarial help. This work was supported by National Institutes of Health Grants AM31400 and AM25532. Cloning was done under P-1/EK-1 conditions of containment in association with National Institutes of Health guidelines involving recombinant DNA molecules.

- Reichlin, S. (1983) *N. Engl. J. Med.* **309**, 1495-1501.
- Reichlin, S. (1983) *N. Engl. J. Med.* **309**, 1556-1563.
- Patel, Y. C., Wheatly, T. & Wing, C. (1981) *Endocrinology* **109**, 1943-1949.
- Trent, D. F. & Weir, G. C. (1981) *Endocrinology* **108**, 2033-2038.
- Gerich, J. E. & Patton, G. S. (1978) *Med. Clin. North Am.* **62**, 375-392.
- Goodman, R. H., Aron, D. C. & Roos, B. A. (1983) *J. Biol. Chem.* **258**, 5570-5573.
- Lechan, R. M., Goodman, R. H., Rosenblatt, M., Reichlin, S. & Habener, J. F. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 2780-2784.
- Benoit, R., Bohlen, P., Ling, N., Briskin, A., Esch, F., Brazeau, P., Ying, S.-Y. & Guilleman, R. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 917-921.
- Funckes, C. L., Minth, C. D., Deschenes, R., Magazin, M., Tavanimi, M. A., Sheets, M., Collier, K., Weith, H. L., Aron, D. C., Roos, B. A. & Dixon, J. (1983) *J. Biol. Chem.* **258**, 8781-8787.
- Shen, L.-P., Pictet, R. L. & Rutter, W. J. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 4575-4579.
- Rigby, P. W. J., Dieckmann, M., Rhodes, C. & Berg, P. (1977) *J. Mol. Biol.* **113**, 237-246.
- Sargent, T. D., Wu, J. R., Sala-Trepst, T. M., Wallace, R. G., Reyes, A. A. & Bonner, J. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 3256-3260.
- Benton, W. D. & Davis, R. W. (1977) *Science* **196**, 180-182.
- Blattner, F. R., Williams, B. G., Blechl, A. E., Deniston-Thompson, K., Fater, H. E., Furlong, L. A., Grunwald, D. S., Keifer, D., Moore, D. D., Sheldon, E. L. & Smithies, O. (1977) *Science* **196**, 161-163.
- Hanahan, D. & Meselson, M. (1980) *Gene* **10**, 63-67.
- Blin, N. & Stafford, D. W. (1976) *Nucleic Acids Res.* **3**, 2303-2308.
- Southern, E. M. (1975) *J. Mol. Biol.* **98**, 503-517.
- Chirgwin, J., Przybyla, A., MacDonald, R. J. & Rutter, W. J. (1979) *Biochemistry* **19**, 5294-5299.
- Weaver, R. F. & Weissman, C. (1979) *Nucleic Acids Res.* **7**, 1175-1185.
- Maxam, A. & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 560-564.
- Aron, D. C., Muszycki, M., Birnbaum, R. S., Sabo, S. W. & Roos, B. A. (1982) *Endocrinology* **109**, 1830-1834.
- Low, M. J., Lechan, R. M., Rosenblatt, M. & Goodman, R. H. (1983) *Endocrinology* **112** Suppl., 151, abstr. 281.
- Benoit, R., Ling, M., Alford, B. & Guilleman, R. (1982) *Biochem. Biophys. Res. Commun.* **107**, 944-950.
- Hobart, P., Crawford, R., Shen, L., Pictet, R. & Rutter, W. (1980) *Nature (London)* **288**, 137-141.
- Noe, B. D. & Spiess, J. (1983) *Endocrinology* **112** Suppl., 85, abstr. 17.
- Uhler, M., Herbert, E., D'Eustachio, P. & Ruddle, F. D. (1983) *J. Biol. Chem.* **258**, 9444-9453.