Proc. Natl. Acad. Sci. USA
Vol. 82, pp. 2320–2324, April 1985
Biochemistry

Identification of cDNA clones encoding a precursor of rat liver cathepsin B

(proteolytic processing/oligonucleotide probes/thiol proteases/lysosomes)

BLANCA SAN SEGUNDO, SHU JIN CHAN, AND DONALD F. STEINER

Department of Biochemistry and Molecular Biology, University of Chicago, Chicago, IL 60637

Contributed by Donald F. Steiner, December 17, 1984

ABSTRACT **Recent studies have suggested that many lysosomal enzymes, including cathepsin B (EC 3.4.22.1), may be synthesized as larger precursors and proteolytically processed to their mature forms. To determine the structure of the primary translation product of cathepsin B, we have screened a phage cDNA library for clones encoding rat liver cathepsin B. We synthesized two extended DNA oligonucleotides to use as hybridization probes: a 50-mer corresponding to the coding segment for residues 215–231 of mature cathepsin B and a 54-mer corresponding to residues 117–134. After screening 600,000 plaques, five clones were obtained that hybridized to the ^{32}P-labeled 50-mer; of these, two (λrCB3 and λrCB5) also reacted with the 54-mer. DNA sequence analysis confirmed that λrCB3 and λrCB5 both encoded rat liver cathepsin B, and the translated sequence is in agreement with the sequence determined [Takio, K., Towatari, T., Katunuma, N., Teller, D. C. & Titani, K. (1983) *Proc. Natl. Acad. Sci. USA* 80, 3666–3670], except for a tryptophan for glycine substitution at residue 78 and the presence of two amino acids at the junction site of the light and heavy chains. Moreover, the DNA sequence reveals an open reading frame extending beyond the 5′ (NH_2 terminus), and the predicted COOH terminus of the coding sequence for the mature protein is extended by six amino acids. These results confirm that the biosynthesis of cathepsin B involves a larger precursor form and demonstrate the effectiveness of long oligonucleotide probes for screening to detect rare cloned mRNAs.**

Cathepsin B is a lysosomal thiol protease that is structurally related to cathepsin H and papain (1). We have recently demonstrated that the mature 31-kDa form of this enzyme in rat islets of Langerhans is immunologically and biochemically closely related to liver cathepsin B and is derived from a larger precursor of ≈43 kDa (2). We also found that a significant fraction of cathepsin B and its precursor forms is located within the insulin secretion granules in addition to the lysosomes in normal islets or in islet tumor cells (3) and that precursor is secreted into the medium along with insulin in response to glucose stimulation (unpublished data). We have suggested that some procathepsin B intermediate forms may be involved in the processing of proinsulin to insulin (4, 5), while the mature forms of the enzyme may play a role in secretory granule turnover (3).

To more fully define the nature of the initial precursor of cathepsin B, to investigate the processing events involved in generating intermediate and mature forms of the enzyme, and to gain more insight into those structural features of the precursor that might be related to the dual targeting of cathepsin B during its biosynthesis, we have cloned the rat liver enzyme. Here we provide a report of the strategy that proved successful and preliminary results on the characterization of two cloned cDNA fragments that encode the mature enzyme and portions of the precursor region.

MATERIALS AND METHODS

Materials. Protected nucleotide monomers and reagents for DNA synthesis were purchased from Applied Biosystems (Foster City, CA). Oligo(dT)-cellulose was Type T-2 from Collaborative Research. [γ-^{32}P]ATP (specific activity, 5000 Ci/mmol; 1 Ci = 37 GBq), [α-^{32}P]dATP, and [α-^{32}P]dCTP (specific activity, 800 Ci/mmol) were from Amersham. T4 polynucleotide kinase and deoxynucleoside triphosphates were from P-L Biochemicals. DNA polymerase (Klenow fragment) and T4 ligase were from Boehringer Mannheim. Restriction enzyme reactions were performed according to the conditions suggested by the manufacturer (New England Biolabs). Nitrocellulose paper was BA 85 (0.45 μm) from Schleicher & Schuell.

Synthesis of Oligonucleotide Probes. DNA oligonucleotides were synthesized using the phosphoramidite methodology (6) on an Applied Biosystems Model 380A synthesizer and were purified by polyacrylamide gel electrophoresis in 7 M urea. Their nucleotide sequence was confirmed by the Banaszuk method (7). The oligonucleotides synthesized were as follows:

A1 30-mer	5′ -G-T-C-G-C-C-A-A-C-T-C-C-T-G-G-A-A-$\frac{C}{T}$-G-T-C-G-A-C-T-G-G-G-G-C- 3′
A2 30-mer	5′ -A-T-$\frac{C}{T}$-T-T-G-A-A-G-A-A-G-C-C-G-T-T-G-T-C-G-C-C-C-C-A-G-T-C-G- 3′
B1 26-mer	5′ -T-G-C-A-C-C-G-G-T-G-A-G-G-G-C-G-A-C-A-C-C-C-C-G-A-A- 3′
B2 21-mer	5′ -A-A-A-A-T-G-T-G-C-G-A-G-G-C-C-G-G-T-T-A-T- 3′
B3 26-mer	5′ -C-A-C-A-T-T-T-T-G-T-T-G-C-A-T-T-T-C-G-G-G-G-T-G-T-C- 3′
B4 13-mer	5′ -A-T-A-A-C-C-G-G-C-C-T-C-G- 3′

The large probes were assembled by annealing overlapping oligonucleotides as shown in Fig. 1 and filling in with *Escherichia coli* DNA polymerase I. The 50-mer was constructed by annealing 0.1 μg of each oligonucleotide (A1 and A2) in 0.1 M NaCl and successively heating at 70°C for 5 min, 37°C for 20 min, 23°C for 20 min, and 4°C for 20 min. The 54-mer was constructed in a two-step process (Fig. 1).

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.

Abbreviation: bp, base pair(s).

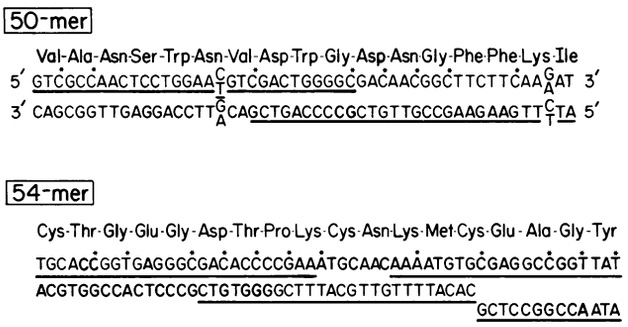

FIG. 1. Synthetic oligonucleotide probes used to screen recombinant phages containing sequences for rat cathepsin B. Sequences underlined are overlapping oligonucleotides that were chemically synthesized. Asterisks denote nucleotides of the synthetic probe that did not match the sequence of the isolated clones.

First, a 39-mer was isolated from an annealed and ligated mixture of oligonucleotides B2, B3, and B4 (8). The 39-mer (0.1 μg) was then annealed to 0.1 μg of the 26-mer (B1) and filled in with DNA polymerase. Reactions were carried out in 50 mM Tris·HCl, pH 7.5/10 mM $MgSO_4$/70 μM dGTP, and dTTP containing 50 μCi of [α-^{32}P]dATP, 50 μCi of [α-^{32}P]dCTP, and 2 units of DNA polymerase (Klenow) in a volume of 27 μl for 2 hr at 15°C. Reactions were chased for a further 15 min by the addition of 2 μl of 5 mM dNTP, and the polymerase was inactivated by heating the reaction mixture at 65°C for 5 min. To increase the specific activity of the probes, oligonucleotides A1, A2, and B4 were first labeled to high specific activity (12×10^6 cpm/pmol) with T4 polynucleotide kinase and [γ-^{32}P]ATP. The oligonucleotide probes were separated from the unincorporated radiolabeled nucleotides by gel filtration on Sephadex G-100 fine equilibrated in 10 mM Tris·HCl, pH 8.0/1 mM EDTA and used directly for hybridization. The specific activity obtained was 5–10 $\times$ 10^8 cpm/μg.

Screening the Library. A rat liver cDNA library constructed in λgt 11 was obtained from R. Hynes and was screened basically as described by Benton and Davis (9). After plating on *E. coli*, LE392, a total of 600,000 plaques were screened on three 0.7% agarose plates (20 $\times$ 30 cm). Filters were pretreated (16 hr) and hybridized (16 hr) in a solution containing 2.5× Denhardt's solution/5× NaCl/Cit/20 mM sodium phosphate, pH 6.5/0.1% $NaDodSO_4$/20% formamide/10 μg of tRNA per ml/50 μg of sonicated denatured salmon sperm DNA per ml at 37°C. (1× Denhardt's solution = 0.02% bovine serum albumin/0.02% Ficoll/0.02% polyvinylpyrrolidone; 1× NaCl/Cit = 0.15 M NaCl/0.015 M Na citrate). Approximately 2.5×10^5 cpm of radioactive DNA probe per ml was used during hybridization. Hybridized filters were washed in succession for 20 min in 2× NaCl/Cit/0.1% $NaDodSO_4$ at room temperature (twice), 1× NaCl/Cit/0.1% $NaDodSO_4$ and 0.5× NaCl/Cit/0.1% $NaDodSO_4$ at 37°C (one time each), and 0.25× NaCl/Cit/0.1% $NaDodSO_4$ at 37°C (twice). The dried filters were autoradiographed at −70°C with Kodak XAR-5 film and DuPont intensifying screens. Positive recombinant clones were isolated by plaque purification, and phage DNA was prepared for further analysis or subcloning (8).

Subcloning. The *Eco*RI fragment from clone λrcB3 was isolated by polyacrylamide gel electrophoresis followed by electroelution and was subcloned into the *Eco*RI site of pUC9 in both orientations. Transformation was done by the calcium chloride method (10). Plasmid DNA was isolated by the cleared lysate procedure (11).

RESULTS

Our approach to the cloning of procathepsin B was through the use of chemically synthesized DNA probes designed on the basis of the amino acid sequence of the mature enzyme (1). A general method using pools of short oligonucleotides that include all possible DNA sequences coding for a given amino acid sequence has been widely used (12). Initially, we also used two short oligonucleotide mixtures, 14 and 17 nucleotides long, corresponding to the predicted coding segment for residues 29–32 (C-C-G/A-A-A-N-G-C-C-C-A-G/A-C-A) and 219–224 (C-C-C-C-A-G/A-T-C-N-A-C-G/A-T-T-C-C-A), respectively. However, positive results were not obtained either via screening cDNA or genomic libraries with these probes, or by using them as specific primers to prepare cDNA libraries. We then turned to the use of extended synthetic oligonucleotides for screening. This method proved successful for detecting cathepsin B clones. A similar strategy was used by Ullrich and co-workers for the isolation of the cDNA encoding the human epidermal growth factor receptor (13) and the gene for human insulin-like growth factor I (14).

In designing the probes, we selected two regions with low codon degeneracy from the central and carboxyl-terminal regions of cathepsin B. The 50-mer corresponded to a region of high homology between cathepsin B and cathepsin H near the carboxyl-terminus, while the 54-mer was chosen from a central region where these two enzymes differ (1). Third position choices were made according to codon usage frequencies in mammalian genes (15). After screening 600,000 plaques, we found 5 positive clones that hybridized to the ^{32}P-labeled 50-mer, but only two of these (λrcB-3 and λrcB-5) also reacted with the ^{32}P-labeled 54-mer. This suggests that the remaining positive clones from the initial screening either may be shorter clones or may correspond to other homologous thiol proteases.

The homology between each synthetic oligonucleotide and the actual cDNA sequence was 77.7% in the case of the 54-mer and 84% in the case of the 50-mer sequence. This indicates that occasional incorrect codon choices do not affect seriously the specificity of the probe and the use of proper hybridization/washing conditions is enough to give very stable signals.

DNA sequence analysis and restriction mapping (Fig. 2) confirmed that both λrcB-3 and λrcB-5 encode rat liver cathepsin B. The sequence of λrcB-3 begins at residue 58 in the amino acid sequence and extends ≈940 base pairs (bp) into the 3′ untranslated region (Fig. 3). A discrepancy from the published amino acid sequence (1) was detected at position 78 where the residue predicted was tryptophan rather than glycine. The nucleotide sequence also discloses the existence of a six-residue carboxyl-terminal extension following the published cathepsin B sequence.

The restriction map of λrcB5 (Fig. 2) shows that it is ≈1960 bp long and that it includes the λrcB3 sequence. The 3′ end was found to contain ≈100 bp of poly(dA), and therefore this clone contains the entire 3′ untranslated region. At the 5′ end the sequence of λrcB-5 extends 360 nucleotides beyond the λrcB3 insert. Preliminary sequence analysis of this region discloses the presence of a contiguous open reading frame extending beyond the amino terminus of mature cathepsin B. This indicates that, as expected, cathepsin B is contained within a larger precursor protein having a relatively large amino-terminal extension in addition to the short carboxyl-terminal extension mentioned above. Furthermore, on RNA blot analysis of rat liver mRNA, the λrcB3 insert specifically hybridizes to an mRNA of 2.3 kilobases (Fig. 4). Thus, the mRNA is ≈30 bp larger than λrcB-5 and would be sufficient to encode a 43-kDa precursor as found in the biosynthetic studies (2). Elucidation of the putative 12-kDa NH_2-terminal extension will require the cloning of a full-length cDNA.

Partial nucleotide sequence analysis of λrcB5 revealed the

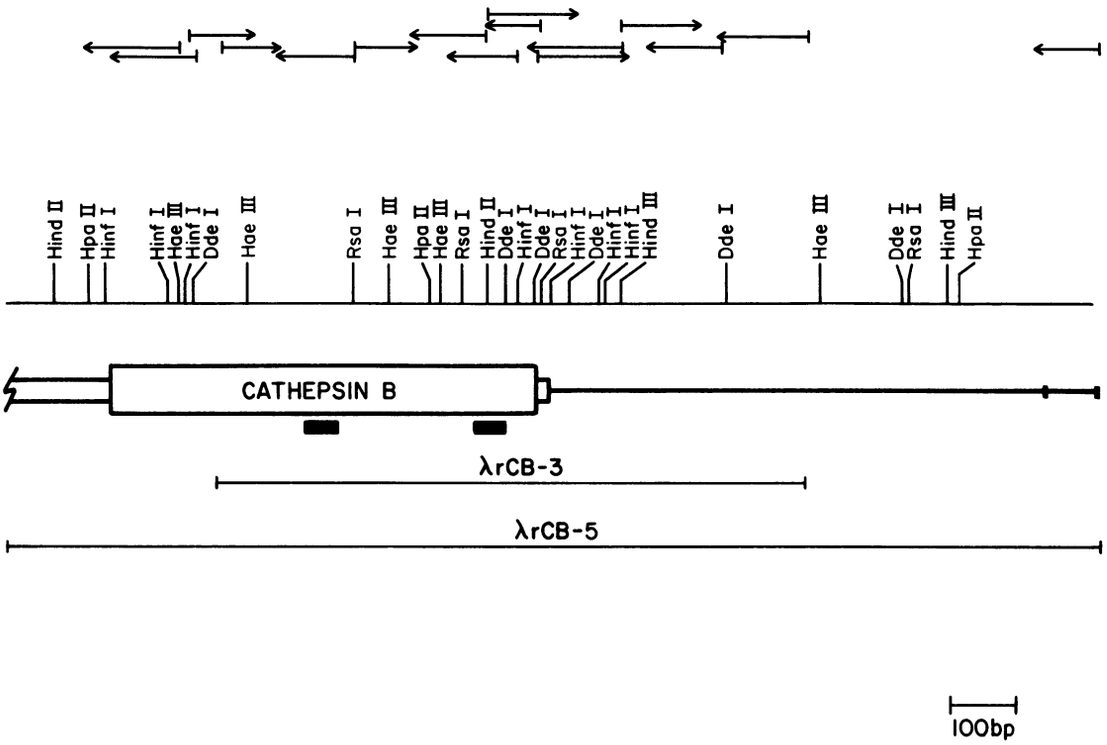

FIG. 2. Restriction map of *Eco*RI fragments of λrCB3 and λrCB5. Lines above the restriction map show the strategy used to determine the nucleotide sequence. Black boxes denote regions used to design the synthetic oligonucleotide probes. Open boxes extending at both ends of cathepsin B correspond to the precursor region. Nucleotide sequence analysis was carried out by the Maxam and Gilbert technique (16) for the *Eco*RI fragment of λrCB3 and subcloned in pUC9 or directly from the *Eco*RI insert of λrCB5 isolated by polyacrylamide gel electrophoreses and electroelution. T-specific reactions were also done by $KMnO_4$ treatment (17).

coding sequence for (pro)cathepsin B in the region from residues −11 through +58 (Fig. 3). In addition to the tryptophan for glycine substitution at position 78, the DNA sequence also reveals that a dipeptide (Gly·Arg) is present between residues 47 and 48—i.e., at the junction between the light and heavy chains—indicating that additional processing occurs after scission of the chain in this region.

DISCUSSION

The mechanisms leading to the specific intracellular targeting of proteins into membranes, lysosomes, or secretory vesicles represents an important unsolved problem in cell biology. In the case of the lysosomal enzymes, specific targeting is believed to be accomplished via at least two pathways. The first, and best defined of these, is the mannose 6-phosphate receptor pathway (20, 21), which operates in the *cis* region of the Golgi apparatus (22, 23) to divert lysosomal enzymes or their precursors bearing phosphorylated mannose-rich oligosaccharide side chains into small vesicles for transport to the lysosomes. Another recognition mechanism that appears to be operative in many eukaryotic cells relies on other, as yet undefined, structural features of these proteins (24). But despite the existence of these mechanisms, it is clear that some lysosomal enzyme precursors normally escape being shunted into the lysosomes and continue along the secretory pathway, entering storage granules and ultimately being secreted (20). Recycling of these secreted forms to the lysosomes can then occur via receptor-mediated endocytosis (21). However, the persistence of a fraction of the lysosomal enzyme precursors in the secretory pathway provides a mechanism whereby these hydrolytic enzymes can be copackaged into secretory granules along with secretory protein precursors and, if suitably activated therein, may participate in the processing of secretory protein precursors such as proinsulin and other prohormones (2, 3).

It is interesting to note that the cathepsin B precursor we have identified appears to be extended at both ends to give the precursor protein estimated on the basis of biosynthetic studies (2) to be ≈43 kDa. It is unclear whether the COOH-terminal extension of six amino acids that we have identified from the cDNA sequence is necessary to modulate the activity of cathepsin B. It seems possible that this sequence, as well as the dipeptide sequence found between the chains, could be removed (viz. in the lysosomes) by stepwise cleavage from the COOH terminus via the known carboxydipeptidase activity of mature cathepsin B (25, 26). It has been proposed (27) that all lysosomal enzymes may be synthesized with a transient COOH-terminal sequence that could function as a sorting sequence, being cleaved after completion of targeting. The structure of the carbohydrate side chain of cathepsin B differs significantly from those found on most of the other lysosomal enzymes (28), which might be in agreement with the above-mentioned hypothesis. Carboxyl-terminal processing of other precursors of lysosomal enzymes has also been noted—e.g., cathepsin D and β glucuronidase (27)—but the exact size and amino acid sequence of these extensions are not yet known.

The processing required to generate the mature (single chain) lysosomal form of cathepsin B appears to involve cleavage at an asparagine residue preceding the NH_2-terminal leucine residue of the light chain. This cleavage also could theoretically be carried out by mature cathepsin B. Further elucidation of the structure of the NH_2-terminal extension should enable the prediction of additional sites of proteolysis necessary to generate the larger 38-kDa form found in both liver and islets (3) and the sites for pepsin cleavage to generate the slightly larger than normal forms of cathepsin B from the precursors secreted by certain tumors (29, 30).

The existence of two forms of cathepsin B in porcine spleen that differ by at least one amino acid within the se-

FIG. 3. Nucleotide sequence of the insert from clone λrCB3 and portions of λrCB5. Predicted amino acid sequence for the mature enzyme and portions of the precursor region are shown above the nucleotide sequences. Residues enclosed in dashed boxes are absent in the mature two-chain enzyme. Circled residue differs from the published amino acid sequence (1).

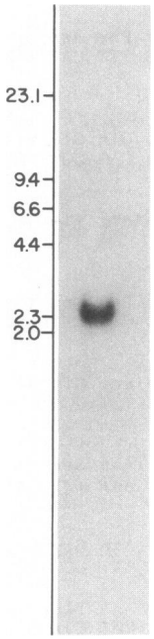

FIG. 4. RNA blot analysis of rat liver mRNA. Numbers on left indicate sizes (in kilobases) for the markers (*Hin*dIII-digested λDNA). Total RNA was extracted from rat liver by a modified guanidine thiocyanate procedure (18). Poly(A) RNA was isolated on oligo(dT)-cellulose. The *Eco*RI insert from λrCB3, subcloned in pUC9, was isolated by polyacrylamide electrophoresis, electroeluted, and nick-translated for use as a hybridization probe. RNA blot analysis was carried out essentially as described by Thomas (19) with minor modifications in washing steps. Poly(A) RNA (10 μg) was glyoxylated and separated on 1% agarose gel. Hybridization was done for about 24 hr at 42°C using 250,000 cpm of the nick-translated probe. RNA blots were washed with four changes of 2× NaCl/Cit/0.1% $NaDodSO_4$ for 30 min each at room temperature and then washed with two changes of 0.1× NaCl/Cit/0.1% $NaDodSO_4$ for 30 min at 50°C. Dried filters were autoradiographed at −70°C with Kodak XAR-5 film and DuPont intensifying screens.

quence near the single NH_2-linked glycosylation site has been reported recently (28). Our sequence for this region agrees with that of the major porcine form (28) and with that of Takio *et al.* (1) except for the presence of tryptophan instead of glycine at position 78. This difference was noted in both of our cDNA clones, making it unlikely to be due to a cloning error. The availability of cathepsin B cDNA probes will allow cloning of other closely related cathepsins (H and L) and papain. These clones will also be useful for the analysis of the expression of these enzymes in different tissues, to investigate the organization of the cathepsin B gene(s) and the role of cathepsin B-like proteases in human tumors in greater detail (31).

The authors are grateful to R. Hynes (Massachusetts Institute of Technology) for kindly providing the rat liver cDNA library used in these studies. We also wish to thank K. Mullis (Cetus Corporation, Emeryville, CA) and A. Efstratiadis (Columbia University) for generous gifts of short oligomers (17-mer and 14-mer, respectively). We are grateful to Michael Welsh and Yi Juan Lu for assistance with various aspects of this work, Albert MacKrell for computer-assisted analysis of DNA sequences, and Cathy Christopherson for assistance in preparing this manuscript. Work from this laboratory is supported by grants from the Public Health Service (AM 13914 and AM 20595).

1. Takio, K., Towatari, T., Katunuma, N., Teller, D. C. & Titani, K. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 3666–3670.
2. Steiner, D. F., Docherty, K. & Carroll, R. (1984) *J. Cell. Biochem.* **24,** 121–130.
3. Docherty, K., Hutton, J. C. & Steiner, D. F. (1984) *J. Biol. Chem.* **259,** 6041–6044.
4. Docherty, K., Carroll, R. & Steiner, D. F. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 4613–4617.
5. Docherty, K., Carroll, R. & Steiner, D. F. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 3245–3249.
6. Beaucage, S. L. & Caruthers, M. H. (1981) *Tetrahedron Lett.* **22,** 1859–1862.
7. Banaszuk, A. M., Deugau, K. V., Sherwood, J., Michalak, M. & Glick, B. R. (1983) *Anal. Biochem.* **128,** 281–286.
8. Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY).
9. Benton, W. D. & Davis, R. W. (1977) *Science* **196,** 180–182.
10. Dagert, M. & Ehrlich, S. D. (1979) *Gene* **6,** 23–28.
11. Bolivar, F. & Backman, K. (1979) *Methods Enzymol.* **68,** 245–267.
12. Wallace, R. B., Johnson, M. J., Hirose, T., Miyake, T., Kawashima, E. H. & Itakura, K. (1981) *Nucleic Acids Res.* **9,** 879–894.
13. Ullrich, A., Coussens, L., Hayflick, J. S., Dull, T. J., Gray, A., Tam, A. W., Lee, J., Yarden, Y., Libermann, T. A., Schlessinger, J., Downward, J., Mayes, E. L. V., Whittle, N., Waterfield, M. D. & Seeburg, P. H. (1984) *Nature (London)* **309,** 418–425.
14. Ullrich, A., Berman, C. H., Dull, T. J., Gray, A. & Less, J. M. (1984) *EMBO J.* **3,** 361–364.
15. Grantham, R., Gantier, C., Gong, M., Jacobzone, M. & Mercier, R. (1981) *Nucleic Acids Res.* **9,** r43–r74.
16. Maxam, A. M. & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 560–564.
17. Rubin, M. & Schmid, C. W. (1980) *Nucleic Acids Res.* **20,** 4613–4618.
18. Feramisco, J. R., Melfman, D. M., Smart, J. E., Burridge, K. & Thomas, G. P. (1982) *J. Biol. Chem.* **257,** 11024–11031.
19. Thomas, P. S. (1980) *Proc. Natl. Acad. Sci. USA* **77,** 5201–5205.
20. Neufeld, E. F. & Ashwell, G. (1980) in *The Biochemistry of Glycoproteins and Proteoglycans*, ed. Lennarz, W. J. (Plenum, New York), pp. 241–266.
21. Sly, W. S. & Fisher, D. (1982) *J. Cell. Biochem.* **18,** 67–85.
22. Pohlmann, R., Waheed, A., Hasilik, A. & von Figura, K. (1982) *J. Biol. Chem.* **257,** 5323–5325.
23. Brown, W. J. & Farquhar, M. G. (1984) *Cell* **36,** 295–307.
24. Stevens, T., Esmon, B. & Schekman, R. (1982) *Cell* **30,** 439–448.
25. Aronson, N. N. & Barrett, A. J. (1978) *Biochem. J.* **171,** 759–765.
26. Bond, S. J. & Barrett, A. J. (1980) *Biochem. J.* **189,** 17–25.
27. Erickson, A. H. & Blobel, G. (1983) *Biochemistry* **22,** 5201–5205.
28. Takahashi, T., Schmidt, P. G. & Tang, J. (1984) *J. Biol. Chem.* **259,** 6059–6062.
29. Recklies, A. D., Mort, J. S. & Poole, A. R. (1982) *Cancer Res.* **42,** 1026–1032.
30. Mort, J. S., Leduc, M. S. & Recklies, A. D. (1983) *Biochim. Biophys. Acta.* **755,** 369–375.
31. Mullins, D. E. & Rohrlich, S. T. (1983) *Biochim. Biophys. Acta* **695,** 177–214.