The EMBO Journal vol.4 no.1 pp.133 – 138, 1985

Genes for the α and γ subunits of mouse nerve growth factor are contiguous

Bronwyn A.Evans[1] and Robert I.Richards[1]

Centre for Recombinant DNA Research and Department of Genetics, Research School of Biological Sciences, Australian National University, Canberra, ACT 2601, Australia

[1]Present address: Howard Florey Institute of Experimental Physiology and Medicine, University of Melbourne, Parkville, Victoria 3052, Australia

Communicated by H.Thoenen

Here we describe the structure and linkage of genes encoding the α and γ subunits of mouse nerve growth factor (NGF). These genes are members of the highly homologous glandular kallikrein multigene family. Together with the β subunit, the α and γ proteins constitute the high mol. wt. (7S) form of NGF isolated from mouse submandibular gland. The γ subunit is an active serine protease and is thought to cleave pro-β-NGF to generate the mature growth factor. The α subunit has no detectable proteolytic activity, but is essential for the stable formation of 7S NGF. Lack of enzyme activity of the α subunit can be attributed, at least in part, to the deletion of 15 nucleotides in a highly conserved coding region which is normally involved in the activation of serine proteases from their inactive zymogen form.

Key words: α and γ NGF genes/DNA sequence/kallikreins/mouse salivary gland

Introduction

Sympathetic and sensory neurons are thought to require nerve growth factor (NGF) for their growth and survival (Bradshaw, 1978; Yanker and Shooter, 1982). NGF has been isolated from the submandibular gland of adult male mice as a high mol. wt. complex (7S NGF) comprising six polypeptides, $\alpha_2\beta_2\gamma_2$, and 1 – 2 g-atoms of Zn^{2+} (Varon *et al.*, 1968; Stach *et al.*, 1980). It is the β chains of 7S NGF which account for its observed growth-promoting activity (Bradshaw, 1978). Both the α and γ subunits have been shown by sequence homology to belong to the family of glandular kallikreins, a subset of the mammalian serine proteases (Ronne *et al.*, 1984; Thomas *et al.*, 1981; Dayhoff, 1978). γ NGF shows the arginylesteropeptidase activity characteristic of kallikreins (Greene *et al.*, 1969), and is believed to cleave pro-β-NGF at two or more sites to generate the active growth factor (Berger and Shooter, 1977; Scott *et al.*, 1983). The α subunit, on the other hand, shows no measurable enzymatic activity, though its presence is apparently necessary for the formation of stable 7S complex (Varon *et al.*, 1968; Greene *et al.*, 1968).

We have undertaken a systematic analysis of the multigene family encoding mouse glandular kallikreins, to assess the possible physiological roles of different members, and to examine mechanisms for the regulation of their tissue-specific expression. Our analysis has shown the gene family to comprise at least 24 non-allelic genes (B.A.Evans and R.I. Richards, in preparation), all of which are located on chromosome 7 (Mason *et al.*, 1983). The finding that most of the genes are closely linked supports the idea that the kallikrein gene family represents a single genetic locus in the mouse. A second important observation is that although different kallikrein genes show a high degree of overall nucleotide sequence homology (80 – 90%), there is marked variability in certain coding regions. This variability is correlated with amino acid residues involved in determining substrate specificity. We have suggested previously, therefore, that members of the kallikrein gene family may have roles in the proteolytic activation of a variety of different growth factors and other peptides (Mason *et al.*, 1983). Here we describe the linkage and structure of two kallikrein genes encoding the α and γ subunits of mouse nerve growth factor.

Results

Identification of linked α and γ NGF genes

Limited nucleotide sequence analysis of a number of kallikrein genes has facilitated the identification of those which encode products already characterized by amino acid sequencing. Using this approach, we have identified the genes mGK-4 and mGK-3 as those encoding the α and γ subunits of NGF, respectively. The two genes are contiguous, transcribed from the same DNA strand, and are separated by 5.3 kb of intergenic DNA (Figure 1). We have not yet isolated any genomic DNA clones which map 5′ to mGK-3. There are, however, two more kallikrein genes, mGK-5 and 6, which are 5.2 and 14.2 kb 3′ to mGK-4, respectively (data not shown). Neither of these genes shows any greater homology to mGK-3 or 4 than that observed between all members of the kallikrein gene family (van Leeuwen and Drinkwater, personal communications).

Nucleotide sequence of the α and γ NGF genes

For both mGK-3 and mGK-4, we have determined the complete nucleotide sequence of coding regions and of introns B, C and D, as well as 400 bp of 5′-flanking DNA (Figure 2). The assignment of coding regions and intervening sequences was made initially by comparison of the two nucleotide sequences with that of mGK-1 (Mason *et al.*, 1983). These assignments were then checked by aligning the translation products of coding regions with the known amino acid sequences of α and γ NGF (Ronne *et al.*, 1984; Isackson and Bradshaw, 1984; Thomas *et al.*, 1981).

The predicted amino acid sequence of the protein encoded by mGK-3 is identical to that determined for γ NGF by Thomas *et al.* (1981). In addition, the nucleotide sequence of the mGK-3 coding regions is virtually identical to that of a partial γ NGF cDNA clone (Howles *et al.*, 1984). The only difference between the two sequences is a T to A change in the cDNA, resulting in glutamate-179 instead of aspartate. It is most likely that this single discrepancy reflects a polymorphism between the different strains of mice used in these studies (BALB/c, DBA/2J and Swiss Webster).

Our initial identification of mGK-4 as the gene encoding α NGF was based on partial amino acid sequences determined by Ronne *et al.* (1984) and by Isackson and Bradshaw (1984).

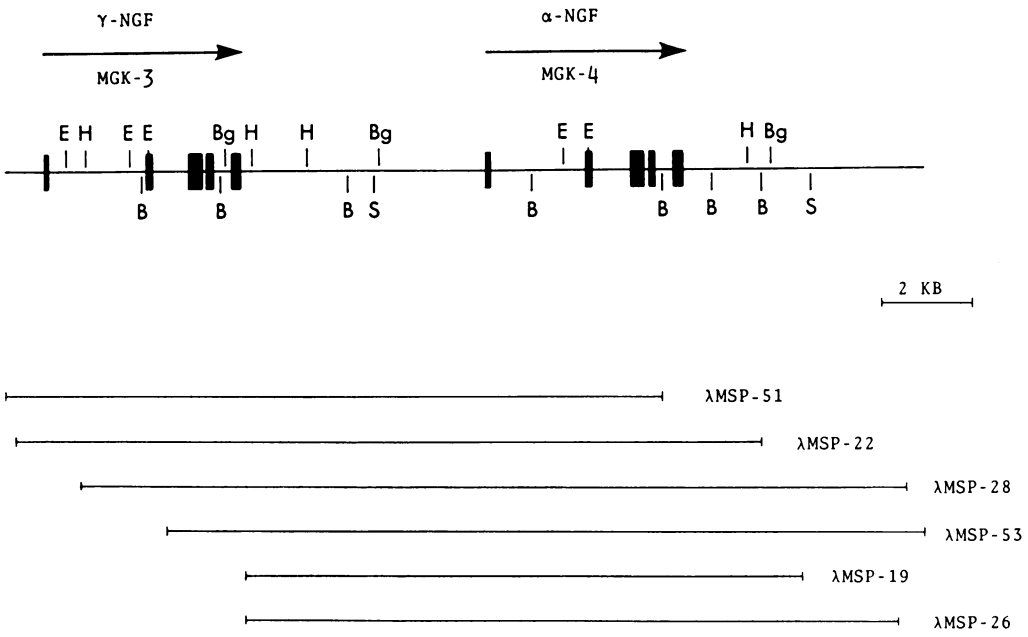


Fig. 1. Restriction map of cloned mouse genomic DNA fragments containing the α and γ NGF genes. Linkage of the genes encoding α and γ NGF is demonstrated by the isolation of a series of overlapping genomic clones (designated λMSP), four of which contain at least part of both mGK-3 and mGK-4. Vertical bars denote the five exons of each gene, while arrows above the map show the direction of transcription. Restriction sites shown are: *Eco*RI (E), *Hind*III (H), *Bam*HI (B), *Bgl*II (Bg), and *Sac*I (S).

As shown in Figure 3, our predicted α NGF protein sequence shows two discrepancies, at positions 25 and 121, with the sequence of Ronne *et al.* (1984). We find, however, that the mGK-4 encoded α NGF has an amino acid sequence identical to those regions determined by Isackson and Bradshaw (1984), and to that predicted from a cDNA clone recently isolated by Isackson *et al.* (1985). There are five silent nucleotide changes between mGK-4 and the α NGF cDNA, again probably reflecting strain differences.

Comparison of the coding regions of mGK-3 and mGK-4 with cDNA clones for γ and α NGF (Howles *et al.*, 1984; Isackson *et al.*, 1985) confirms that these genes have exon-intron boundaries which are identical to those of mGK-1 (Mason *et al.*, 1983).

Structure of the α NGF protein

The DNA sequence of mGK-4 indicates that the α NGF protein has an altered N terminus (Ronne *et al.*, 1984) due to small deletions which result in the loss of 15 nucleotides from exon 2 (Figure 4). The simplest explanation for the observed mGK-4 sequence is a series of five deletions, varying from one to eight nucleotides. Due to sequence redundancy in this region, however, there are at least 120 different combinations of deletion events which could give rise to the observed outcome.

No other mutation events in mGK-4 affect the α NGF protein to the same extent as those in exon 2. Most of the amino acid differences between, for example, the α and γ subunits of NGF, occur in regions which are characteristically variable amongst members of the kallikrein family (Figure 3). The principal exceptions to this observation are a lysine to serine change at position 182, an aspartate to tyrosine change at position 183, and a glycine to histidine change at position 187. All of these residues are in close proximity to the serine-189 which participates in formation of the active site (Dayhoff, 1978; Krieger *et al.*, 1974).

Comparison of our sequence with the protein (Ronne *et al.*, 1984; Isackson and Bradshaw, 1984) and cDNA (Isackson *et al.*, 1985) sequences indicates that the mGK-4 transcript is initiated, terminated and processed correctly.

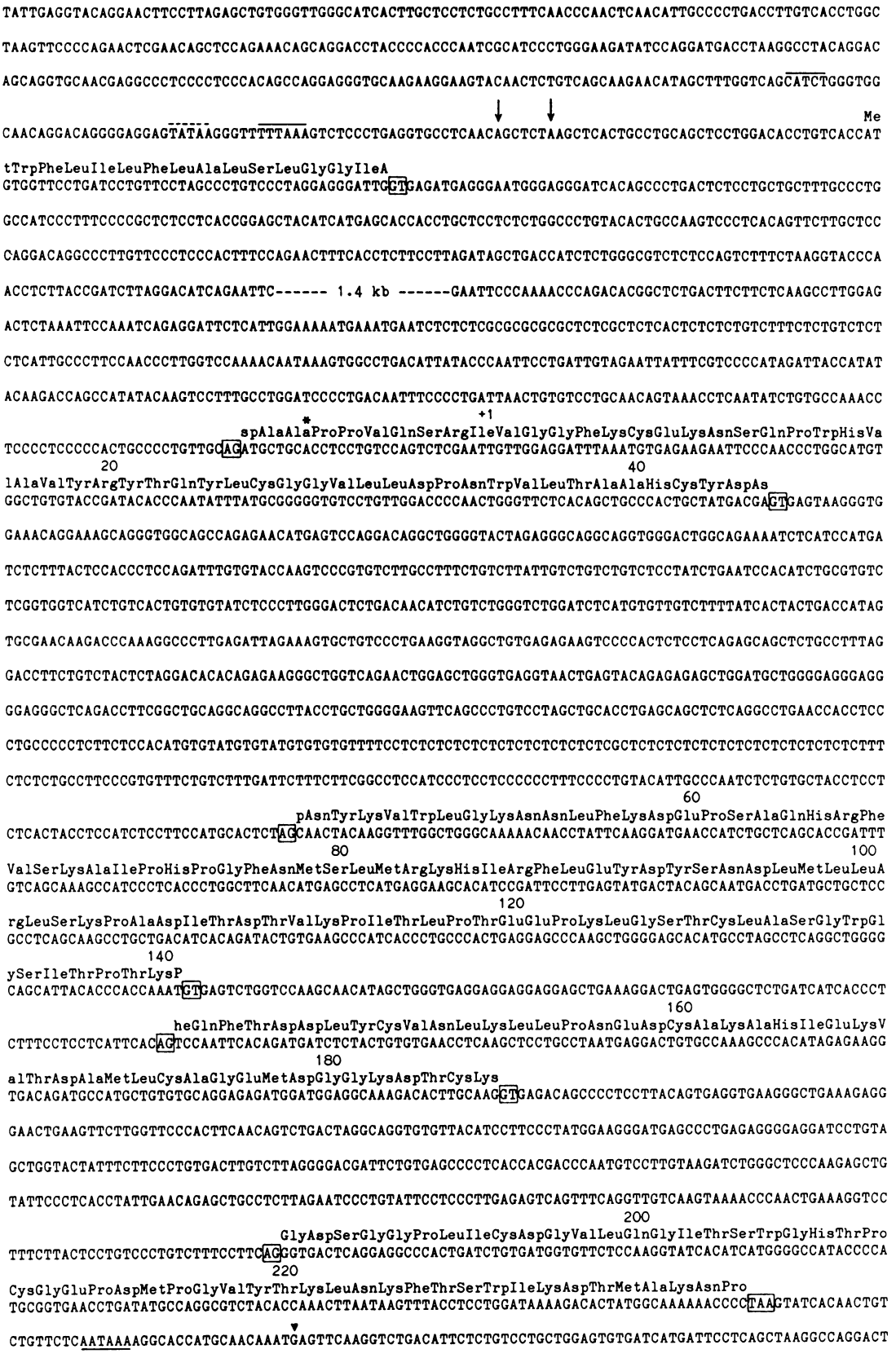
Discussion

Sequencing of the α and γ NGF genes allows us to predict the complete amino acid sequence of the precursor forms of these subunits. In conjunction with the analysis of cDNA clones for mouse prepro-β NGF (Scott *et al.*, 1983; Ullrich *et al.*, 1983), our data provide a structural description of all the components of 7S NGF.

The relationship between the amino acid sequences of the α and γ subunits of NGF is of particular interest. Both of these subunits bind specifically to β NGF, albeit at different sites (Silverman and Bradshaw, 1982), but only the γ subunit shows any catalytic activity (Varon *et al.*, 1968). We have shown that a series of deletions in exon 2 of mGK-4 result in profound changes in the amino acid sequence of the N terminus of α NGF. One consequence of the deletions is that glutamine replaces the arginine (position −1, Figure 3) required for cleavage of the zymogen peptide. It is thought that removal of this peptide is accompanied by a conformational change necessary for serine protease activation (Bode, 1983). In addition, the amino acids normally at the N terminus of the mature enzyme, namely Ile-Ile, Ile-Val or Val-Val, have been implicated in stabilization of the active site (Bode, 1979). In the case of α NGF, the usual N-terminal isoleucine/valine residue is absent.

The precise cause of lack of enzyme activity cannot be assigned at this stage to any particular mutation, since additional changes affect residues close to the serine-189 involved in active site formation. The deletion events in exon 2 are unique, however, compared with other members of the kallikrein gene family (Mason *et al.*, 1983, and unpublished observations). Other amino acid changes in the coding region are the result of point mutations, a common mechanism in the varia-

a


TATTGAGGTACAGGAACTTCCTTAGAGCTGTGGGTTGGGCATCACTTGCTCCTCTGCCTTTCAACCCAACTCAACATTGCCCCTGACCTTGTCACCTGGC 100
TAAGTTCCCCAGAACTCGAACAGCTCCAGAAACAGCAGGACCTACCCCACCCAATCGCATCCCTGGGAAGATATCCAGGATGACCTAAGGCCTACAGGAC 200
AGCAGGTGCAACGAGGCCCTCCCCTCCCACAGCCAGGAGGGTGCAAGAAGGAAGTACAACTCTGTCAGCAAGAACATAGCTTTGGTCAGCATCTGGGTGG 300
Me
CAACAGGACAGGGGAGGAGTATAAGGGTTTTTAAAGTCTCCCTGAGGTGCCTCAACAGCTCTAAGCTCACTGCCTGCAGCTCCTGGACACCTGTCACCAT 400
tTrpPheLeuIleLeuPheLeuAlaLeuSerLeuGlyGlyIleA
GTGGTTCCTGATCCTGTTCCTAGCCCTGTCCCTAGGAGGGATTGGTGAGATGAGGGAATGGGAGGGATCACAGCCCTGACTCTCCTGCTGCTTTGCCCTG 500
GCCATCCCTTTCCCCGCTCTCCTCACCGGAGCTACATCATGAGCACCACCTGCTCCTCTCTGGCCCTGTACACTGCCAAGTCCCTCACAGTTCTTGCTCC 600
CAGGACAGGCCCTTGTTCCCTCCCACTTTCCAGAACTTTCACCTCTTCCTTAGATAGCTGACCATCTCTGGGCGTCTCTCCAGTCTTTCTAAGGTACCCA 700
ACCTCTTACCGATCTTAGGACATCAGAATTC------ 1.4 kb ------GAATTCCCAAAACCCAGACACGGCTCTGACTTCTTCTCAAGCCTTGGAG 800
ACTCTAAATTCCAAATCAGAGGATTCTCATTGGAAAAATGAAATGAATCTCTCTCGCGCGCGCGCTCTCGCTCTCACTCTCTCTGTCTTTCTCTGTCTCT 900
CTCATTGCCCTTCCAACCCTTGGTCCAAAACAATAAAGTGGCCTGACATTATACCCAATTCCTGATTGTAGAATTATTTCGTCCCCATAGATTACCATAT 1000
ACAAGACCAGCCATATACAAGTCCTTTGCCTGGATCCCCTGACAATTTCCCCTGATTAACTGTGTCCTGCAACAGTAAACCTCAATATCTGTGCCAAACC 1100
+1
spAlaAlaProProValGlnSerArgIleValGlyGlyPheLysCysGluLysAsnSerGlnProTrpHisVa
TCCCCTCCCCCACTGCCCCTGTTGCAGATGCTGCACCTCCTGTCCAGTCTCGAATTGTTGGAGGATTTAAATGTGAGAAGAATTCCCAACCCTGGCATGT 1200
20 40
lAlaValTyrArgTyrThrGlnTyrLeuCysGlyGlyValLeuLeuAspProAsnTrpValLeuThrAlaAlaHisCysTyrAspAs
GGCTGTGTACCGATACACCCAATATTTATGCGGGGGTGTCCTGTTGGACCCCAACTGGGTTCTCACAGCTGCCCACTGCTATGACGAGTGAGTAAGGGTG 1300
GAAACAGGAAAGCAGGGTGGCAGCCAGAGAACATGAGTCCAGGACAGGCTGGGGTACTAGAGGGCAGGCAGGTGGGACTGGCAGAAAATCTCATCCATGA 1400
TCTCTTTACTCCACCCTCCAGATTTGTGTACCAAGTCCCGTGTCTTGCCTTTCTGTCTTATTGTCTGTCTGTCTCCTATCTGAATCCACATCTGCGTGTC 1500
TCGGTGGTCATCTGTCACTGTGTGTATCTCCCTTGGGACTCTGACAACATCTGTCTGGGTCTGGATCTCATGTGTTGTCTTTTATCACTACTGACCATAG 1600
TGCGAACAAGACCCAAAGGCCCTTGAGATTAGAAAGTGCTGTCCCTGAAGGTAGGCTGTGAGAGAAGTCCCCACTCTCCTCAGAGCAGCTCTGCCTTTAG 1700
GACCTTCTGTCTACTCTAGGACACACAGAGAAGGGCTGGTCAGAACTGGAGCTGGGTGAGGTAACTGAGTACAGAGAGAGCTGGATGCTGGGGAGGGAGG 1800
GGAGGGCTCAGACCTTCGGCTGCAGGCAGGCCTTACCTGCTGGGGAAGTTCAGCCCTGTCCTAGCTGCACCTGAGCAGCTCTCAGGCCTGAACCACCTCC 1900
CTGCCCCCTCTTCTCCACATGTGTATGTGTATGTGTGTGTTTTCCTCTCTCTCTCTCTCTCTCTCTCTCGCTCTCTCTCTCTCTCTCTCTCTCTTT 2000
CTCTCTGCCTTCCCGTGTTTCTGTCTTTGATTCTTTCTTCGGCCTCCATCCCTCCTCCCCCCTTTCCCCTGTACATTGCCCAATCTCTGTGCTACCTCCT 2100
60
pAsnTyrLysValTrpLeuGlyLysAsnAsnLeuPheLysAspGluProSerAlaGlnHisArgPhe
CTCACTACCTCCATCTCCTTCCATGCACTCTAGCAACTACAAGGTTTGGCTGGGCAAAAACAACCTATTCAAGGATGAACCATCTGCTCAGCACCGATTT 2200
80 100
ValSerLysAlaIleProHisProGlyPheAsnMetSerLeuMetArgLysHisIleArgPheLeuGluTyrAspTyrSerAsnAspLeuMetLeuLeuA
GTCAGCAAAGCCATCCCTCACCCTGGCTTCAACATGAGCCTCATGAGGAAGCACATCCGATTCCTTGAGTATGACTACAGCAATGACCTGATGCTGCTCC 2300
120
rgLeuSerLysProAlaAspIleThrAspThrValLysProIleThrLeuProThrGluGluProLysLeuGlySerThrCysLeuAlaSerGlyTrpGl
GCCTCAGCAAGCCTGCTGACATCACAGATACTGTGAAGCCCATCACCCTGCCCACTGAGGAGCCCAAGCTGGGGAGCACATGCCTAGCCTCAGGCTGGGG 2400
140
ySerIleThrProThrLysP
CAGCATTACACCCACCAAATGTGAGTCTGGTCCAAGCAACATAGCTGGGTGAGGAGGAGGAGGAGCTGAAAGGACTGAGTGGGGCTCTGATCATCACCCT 2500
160
heGlnPheThrAspAspLeuTyrCysValAsnLeuLysLeuLeuProAsnGluAspCysAlaLysAlaHisIleGluLysV
CTTTCCTCCTCATTCACAGTCCAATTCACAGATGATCTCTACTGTGTGAACCTCAAGCTCCTGCCTAATGAGGACTGTGCCAAAGCCCACATAGAGAAGG 2600
180
alThrAspAlaMetLeuCysAlaGlyGluMetAspGlyGlyLysAspThrCysLys
TGACAGATGCCATGCTGTGTGCAGGAGAGATGGATGGAGGCAAAGACACTTGCAAGGTGAGACAGCCCCTCCTTACAGTGAGGTGAAGGGCTGAAAGAGG 2700
GAACTGAAGTTCTTGGTTCCCACTTCAACAGTCTGACTAGGCAGGTGTGTTACATCCTTCCCTATGGAAGGGATGAGCCCTGAGAGGGGAGGATCCTGTA 2800
GCTGGTACTATTTCTTCCCTGTGACTTGTCTTAGGGGACGATTCTGTGAGCCCCTCACCACGACCCAATGTCCTTGTAAGATCTGGGCTCCCAAGAGCTG 2900
TATTCCCTCACCTATTGAACAGAGCTGCCTCTTAGAATCCCTGTATTCCTCCCTTGAGAGTCAGTTTCAGGTTGTCAAGTAAAACCCAACTGAAAGGTCC 3000
200
GlyAspSerGlyGlyProLeuIleCysAspGlyValLeuGlnGlyIleThrSerTrpGlyHisThrPro
TTTCTTACTCCTGTCCCTGTCTTTCCTTCAGGGTGACTCAGGAGGCCCACTGATCTGTGATGGTGTTCTCCAAGGTATCACATCATGGGGCCATACCCCA 3100
220
CysGlyGluProAspMetProGlyValTyrThrLysLeuAsnLysPheThrSerTrpIleLysAspThrMetAlaLysAsnPro
TGCGGTGAACCTGATATGCCAGGCGTCTACACCAAACTTAATAAGTTTACCTCCTGGATAAAAGACACTATGGCAAAAAACCCCTAAGTATCACAACTGT 3200
CTGTTCTCAATAAAAGGCACCATGCAACAAATGAGTTCAAGGTCTGACATTCTCTGTCCTGCTGGAGTGTGATCATGATTCCTCAGCTAAGGCCAGGACT 3300
TCCAAAGTAGATCAGCAAGGGACGGGTGTAC

b

```
TATTGAGGTGCAGGACCTTATTTAGAGCTGTGAGTTGGGGGTCACTCGCTCCTCCATCCTCCTTTCACACTTCCTGACCATCACCCCTGACCTTGCCATC   100

TGGCTACAGTCCCCGGAACTCAAACAGCTCCAGAAACAGCAGGTCCTGCCCCAACCAATCTCATCCCTGGGAAGGTGTCCAGGGCGACCTAGGGCCTACT   200

GTAGAGCAGGTGCAACGAGGCTCTCCCATCCCACAGCCAGGAGGGTGCAAGCAGGAAGTACAACTCTGTCAGCACGAACAGAGCTTTGGTCAGCATCTGA   300
                                                                                        ↓    ↓
GTGGAAAGGACAGGGGAGGGGCTGTGGGGAGAATGAGGGTTTTAAAGTCTCCCTGAGGAGCCTCCACAGCTCCAAGCTCACTGCCTGCAGCTCCTGGACA   400

              MetTrpPheLeuIleLeuPheLeuAlaLeuSerLeuGlyGlyIleA
CCTGTTACCATGTGGTTCCTGATCCTGTTCCTAGCCCTGTCCCTAGGAGGGATTGGTGAGATGAGGGAAAGGGAGGGGTCACAGCCCTGACTCTCCTGCT   500

GCCTTTGGCC------ 1.8 kb ------CGCGCACACACACACACACACACACACTCACACTACCCTGCCATCCCTTTGCCCAAATCTATAAAGTG   600

ACCTGACACTATACCTGATACTTGATTAAAGAGTTCTTTTGTATCCACAGCATTGATCTCCCTATCGAGGACCAGCATATATGGGTCCCTTCCCCAGATG   700

TCCTCGCCAAGTCCCCTGATCCCAGCCCTGCTTAAATGTGTCCTCCAACAATACACCTCCATATCTTGCCAACTCTCCCCTCCCCGATTGTTTCTGTTGC   800
   *                                                    20
 spAlaAlaProProValGlnSerGlnValAspCysGluAsnSerGlnProTrpHisValAlaValTyrArgPheAsnLysTyrGlnCysGlyGlyVal
AGATGCTGCACCTCCTGTCCAGTCTCAAGTGGACTGTGAGAATTCCCAACCCTGGCATGTGGCTGTGTACCGCTTCAACAAATATCAATGCGGGGGAGTC   900
                       40
LeuLeuAspArgAsnTrpValLeuThrAlaAlaHisCysTyrAsnAs
CTGTTGGACCGCAACTGGGTTCTCACAGCTGCCCACTGCTATAACGAGTGAGTAAGGGTGGAGACAGGAAAGCAGGGTGGCAGCCAGAGATCATGACTCC  1000

AGGACAGGCTGGGGTACTGAAGGGCAGGCAGGTGGAACTGGCTGAAACTCTATCCATGACCTCCTGGGTCTCTCTTTACTCCACCCTCCAGACTGGGGTC  1100

TGAGTACTGTTTCTTTTCTGTCTTATTGTCTGTCTATCTCCTGTCTGTATTCACATCTGAGTGTCTCTGTGGTCATCTGTCACTGTGTGTCTCCCCTTGC  1200

GACTCTGCAAAGCCCTGTCTGGGTCTGGGTCTGGGTCTCATGTGTTCTCTTTTATCACTACTGACCATAGTGGGACCAGGCTCCCTGAGGATAGAACGTG  1300

CTGTCCCTGAGGGTAGAGTGTGAGAGAAGTCCTCACTCTCCTCAGAGCAGCTCTGCCTTTAGGACCTTCTGTCCACTCTGAGGACACACAGAGAAGGGCT  1400

GGTCAGAACTGGAGCTGGGTGAGGTAACTGAGTACAGAGAGAGCTGGATGCTGGGAGATGAGTCGAGAGAGTTCCGACCTTGGCTGCAGGCAGGCCTTAC  1500

CTACTGGGGAAGTTCACCTGAGAAGCTCTCAGGCCTGCCCGTCCTCCCTGCCCCCTCTTCCCTTGTCTTATGTTCCTGTCTCTGTCTCTTTCTGCCTCTC  1600

TGACTCAATATGTATGTATGTATGTATGTATGTATGTATGTATGTATGTATGTATGTATGTGCCTTTCTGTCCCTCTGTGTGTGTTTCTGTCCCGGTTTGTATA  1700

CGTGTATAGGTGTGTGTGTGTCGTTCTCTCTGTGTGTGTTTCTGTCTCTGACTCTTCCTTCCTCACTTCCTCCTCACCTCTTTCCCTGTACTCTGACCAA  1800
                                                                                      60
                                                      pLysTyrGlnValTrpLeuGlyLysAsnAsnPheLeuGluAspGluPro
TCTCCTAGACTCTCTCCTCTCACTCCCTCTGTCTGCTTCCCTGACCCCCAGCAAGTACCAGGTTTGGCTGGGCAAAAACAACTTTTTGGAGGATGAACCT  1900
                                                             80
SerAspGlnHisArgLeuValSerLysAlaIleProHisProAspPheAsnMetSerLeuLeuAsnGluHisThrProGlnProGluAspAspTyrSerA
TCTGACCAACACCGGCTTGTCAGCAAAGCCATCCCTCACCCTGACTTCAACATGAGCCTCCTGAATGAGCACACCCCACAACCTGAGGACGACTACAGCA  2000
               100                                                   120
snAspLeuMetLeuLeuArgLeuSerLysProAlaAspIleThrAspValValLysProIleThrLeuProThrGluGluProLysLeuGlySerThrCy
ATGACCTGATGCTGCTCCGCCTCAGCAAGCCTGCTGACATCACAGATGTTGTGAAGCCCATCACCCTGCCCACTGAGGAGCCCAAGCTGGGGAGCACATG  2100
                          140
sLeuAlaSerGlyTrpGlySerThrThrProIleLysP
CCTAGCCTCAGGCTGGGGCAGCACTACACCCATCAAGTGTGAGTCTTCTCCAAGCAACAGAGAGGGGTGGCAGTGAGGGGCAGAGGGGACTGAGCTGGCC  2200
                                                                                            160
                                         heLysTyrProAspAspLeuGlnCysValAsnLeuLysLeuLeuProAsnGluAspCysAspLysA
CCTGCTCACCACCCTCTTGCCTCCTGATTCACAGTCAAATATCCAGATGATCTCCAGTGTGTGAACCTCAAGCTCCTGCCTAATGAGGACTGTGACAAAG  2300
                                                             180
laHisGluMetLysValThrAspAlaMetLeuCysAlaGlyGluMetAspGlyGlySerTyrThrCysGlu
CACATGAAATGAAGGTGACAGATGCCATGCTGTGTGCAGGAGAGATGGATGGTGGCTCATACACTTGTGAGGTAAGACAGCCCTTCCTTACAGTGAGGTG  2400

AAGGGCTGAAAGAGAGAACTGAGGTTCTTGGTTTCCACTTCAACACCCTGACTAGGCAGGCTCTTTGCATACTTCCCTAAGGAAGGGATGCACCCTGGAA  2500

GGGGAGGATCCTGTAGCTGGTACTATTTCTCCTCTGTGACTTGGCTTAGGGGACTGTCCAGTCCACCCCTCACCCCAACCACATGTCCCTGGAACAGATG  2600

GTCTCCTAAGAGCTGGATTCCCTCACCTACTGGCCAGAGCTACCTCTTAGAGTCCCTGTATTCCTTCCCTGGGAGTCAGTTCCAGGTTGTCAAGTAGAAT  2700
                                                                              200
                                                     HisAspSerGlyGlyProLeuIleCysAspGlyIleLeuGlnGlyIleThrSerT
CCAACCGCCAGTCCTGTCTTACTGCTGTCCCTGTCCTCTTTCAGCATGACTCAGGAGGCCCACTGATCTGTGATGGTATTCTCCAAGGTATCACATCAT  2800
                                                     220
rpGlyProGluProCysGlyGluProThrGluProSerValTyrThrLysLeuIleLysPheSerSerTrpIleArgGluThrMetAlaAsnAsnPro
GGGGCCCTGAACCATGTGGTGAGCCCACTGAGCCAAGTGTCTACACCAAACTTATTAAGTTCAGCTCCTGGATAAGAGAAACTATGGCGAATAACCCCTG  2900
                                              ▼
AGTGTCACATTGTCCCCTTTTCTCAATAAAGCCCACCATGCAGCAAATGAGTCCAAGTTCTGACATTCTCTGGCCTTCTGCTGTCGCAGTACCCAGTCAT  3000

GGTTCACTCGGCTAAGGTCAGGACTTCCAAAGTAGATC
```

Fig. 2. Nucleotide sequence of the α and γ NGF genes. The sequence of the transcribed DNA strand of the γ NGF (**a**, mGK-3) and α NGF (**b**, mGK-4) genes is shown. Figures at the right of each section refer to nucleotide numbers. The splice sites around each exon, and the termination codons are boxed. Numbers above both amino acid sequences refer to the position relative to the N-terminal isoleucine of mature γ NGF (see Figure 3). Asterisks denote the first residue of the zymogen sequence. Amino acids preceding this residue constitute the signal peptide. The CAT boxes (Benoist *et al.*, 1980) and the variant 5′-TTTAAA-3′ boxes (Corden *et al.*, 1980; common to mGK-1, Mason *et al.*, 1983) preceding mGK-3 and mGK-4 are overlined. An alternative 5′-TATAA-3′ box preceding mGK-3 is also marked. Putative transcription initiation sites are shown by arrows (Mason *et al.*, 1983). The polyadenylation signals (Proudfoot and Brownlee, 1976) in the 3′-untranslated regions of mGK-3 and 4 are underlined, and triangles indicate transcription termination sites (Mason *et al.*, 1983).

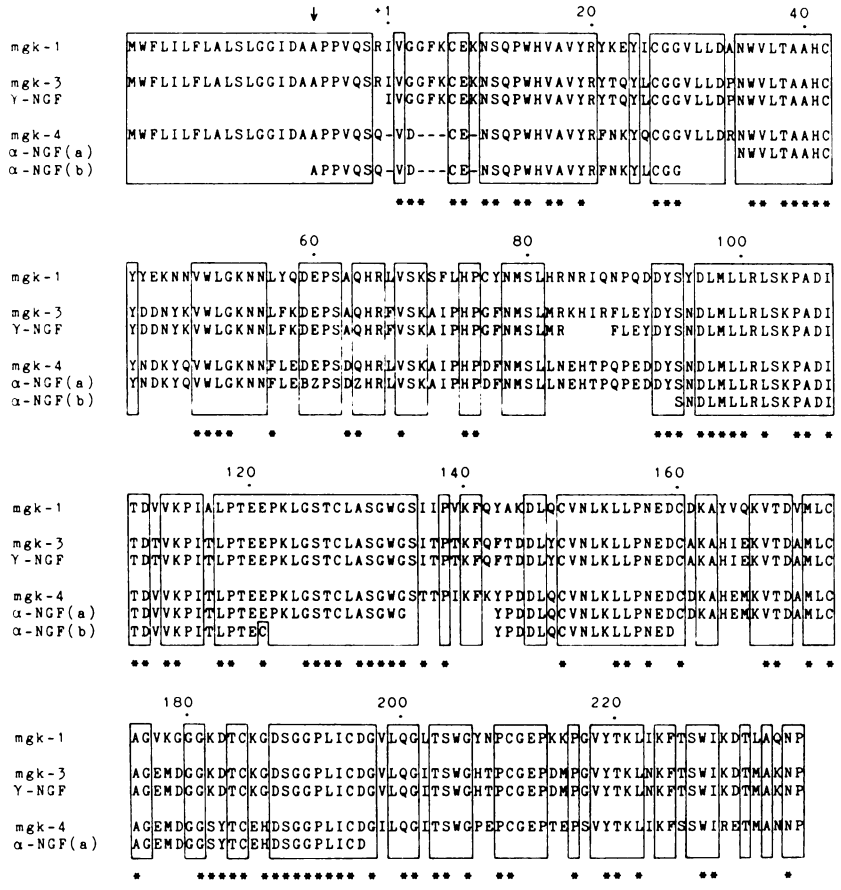


Fig. 3. Comparison of the amino acid sequences of mGK-1, mGK-3 and mGK-4 gene products with those of α and γ NGF. The γ NGF sequence is that determined by Thomas *et al.* (1981). Partial α NGF sequences are those of **(a)** Isackson and Bradshaw (1984) and **(b)** Ronne *et al.* (1984). Boxed areas represent amino acid residues which are identical in mGK-1 (Mason *et al.*, 1983), mGK-3/γ NGF and mGK-4/α NGF. Numbers refer to the position relative to the N-terminal isoleucine of mature γ NGF, and are consistent with those in Figure 2. The arrow represents the first residue of the zymogen peptide or, in the case of α NGF, the N-terminal residue of the mature protein (see text). Asterisks below each set of sequences indicate residues which are identical in all the following members of the glandular kallikrein family: mouse γ NGF, EGF-binding protein type B (Ronne *et al.*, 1983) and γ renin (Poe *et al.*, 1983), rat pancreatic and submaxillary gland kallikrein (Swift *et al.*, 1982), porcine pancreatic kallikrein (Tschesche *et al.*, 1979), and the proteins encoded by pMK-1 (Richards *et al.*, 1982), mGK-1 and mGK-2 (Mason *et al.*, 1983). The residues necessary in formation of the serine protease active site are histidine-41, aspartate-96 and serine-189 (Dayhoff, 1978; Krieger *et al.*, 1974).

tion between different kallikreins (Figure 3). We think it likely therefore that the exon 2 deletion events preceded the point mutations, which could subsequently occur in conserved regions because of the prior loss of the functional constraint of enzyme activity.

We have suggested that differences in the amino acid sequences of members of the kallikrein family are functionally related to their substrate specificity (Mason *et al.*, 1983). In the case of the α and γ subunits, divergence of the two adjacent genes has resulted in recognition of a common substrate but at different sites. In addition, one gene has lost its ability to encode a functional serine protease. The proximity of the two genes and their complementary roles in NGF biosynthesis support the notion that they evolved by a duplication event from a common ancestor itself involved in this process. The availability of the cloned α and γ subunit genes should facilitate an analysis of their respective roles in β NGF biosynthesis. Substitution of the mutated region (exon 2) of the α gene with the corresponding region from the γ gene, and subsequent expression in heterologous cells (Karin *et al.*, 1983; Richards *et al.*, 1984), would allow us to determine whether the resulting gene product has enzyme activity. If this were the case, newly acquired cleavage sites for the chimeric product may reflect the binding sites of the original α subunit on β NGF (Scott *et al.*, 1983).

In addition to functional studies on NGF biosynthesis, we are interested in examining the expression of the α and γ subunits, since both are found in equimolar amounts in the 7S complex of mouse salivary gland. The tight linkage of the mGK-3 and mGK-4 genes raises the possibility of coordinate control of transcription through a single *cis*-acting element, since no other kallikrein gene with a different expression phenotype separates these two members of the gene family. Examination of this possibility will require gene transfer experiments into cell lines expressing these genes (Walker *et al.*, 1983). Alternatively, a *trans*-acting factor may regulate the expression of both genes by binding to a regulatory sequence (Karin *et al.*, 1984) common to the α and γ subunit genes but not to other members of the kallikrein gene family. A comparison of the 5′-flanking regions of mouse glandular kallikrein genes reveals a high degree of homology (>75%) extending >400 bp from the start of transcription (Figure 2, Mason *et al.*, 1983, and unpublished results). The overall homology between the 5′-flanking regions of the α and γ subunit genes is somewhat higher (85%), with one particular region (nucleotides −145 to −95, Figure 2) having 96%

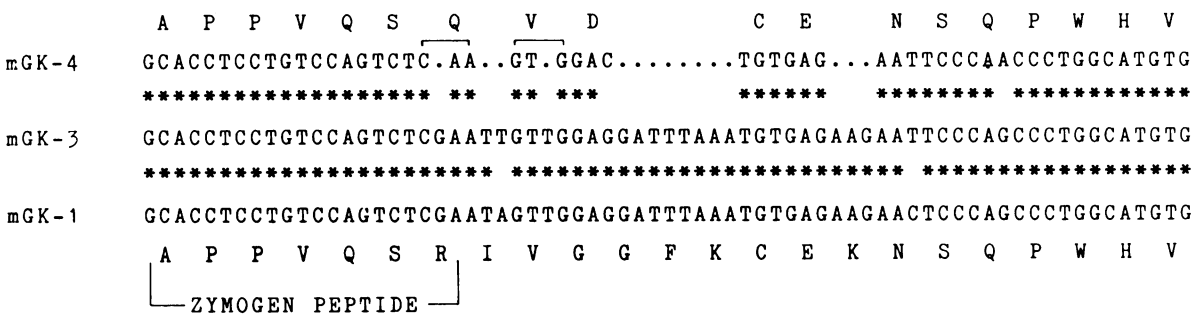


Fig. 4. Nucleotide deletions in mGK-4 affecting the N-terminal structure of α NGF. Partial exon 2 nucleotide sequences are aligned to maximize homology (denoted by asterisks) between mGK-4 and mGK-3 or mGK-1 (Mason *et al.*, 1983). This alignment results in only two single base pair mismatches between mGK-4 and mGK-3, comparable with the observed differences between mGK-3 and mGK-1.

homology (in contrast to 78% or 76% if either of these respective sequences from mGK-3 or mGK-4 are compared with that of mGK-1; Mason *et al.*, 1983). While the greater general degree of flanking region homology may be due to recent duplication of the ancestor to the α and γ subunit genes, the localized conservation of DNA sequence (-145 to -95) may reflect a common regulatory role in the coordinate expression of these two genes.

Finally, the nucleotide sequence analysis of the α and γ subunit genes allows us to design oligodeoxynucleotides complementary to the regions which are most variable relative to other members of the kallikrein gene family. These oligodeoxynucleotides then represent 'gene-specific' probes which can be used to examine tissue-specific expression of α and γ NGF.

Materials and methods

Mouse genomic DNA libraries

Two different mouse genomic libraries were screened by the procedure of Benton and Davis (1977), using the kallikrein cDNA, pMK-1 (Richards *et al.*, 1982), as a hybridization probe. An embryonic BALB/c mouse DNA library, in the bacteriophage vector Charon 28, was the gift of Dr. P.Leder. The clones λMSP-19, 22 and 26 (Figure 1) were isolated from this library. We constructed a second BALB/c library using mouse liver DNA subjected to limited *Sau*3A digestion. This DNA was treated with bacterial alkaline phosphatase (65°C for 10 min) and fractionated over a 5–24% NaCl gradient (37 000 r.p.m. for 4.5 h at 25°C, in an SW41 rotor). DNA fragments of 15–22 kb were then inserted into the EMBL3A vector (Frischauf *et al.*, 1983) and packaged *in vitro* (Enquist and Sternberg, 1979). Screening of an unamplified library of 250 000 phage with pMK-1 gave a total of 45 positives including λMSP-28, 51 and 53. All the positives were analysed by digestion of the DNA with a panel of restriction enzymes (*Eco*RI, *Hind*III, *Bam*HI, *Bgl*II, *Sac*I, *Sal*I and *Sma*I), followed by Southern blotting (Southern, 1975) and hybridization with a series of exon-specific probes (Mason *et al.*, 1983).

Nucleotide sequencing

The nucleotide sequence of the γ NGF (mGK-3) and α NGF (mGK-4) genes was obtained from a series of overlapping subclones in the plasmid pUC13 (Messing, 1983; see Figure 1). These were used as a source of DNA restriction fragments for cloning into M13.mp10 or mp11 (Messing, 1983), and sequencing by the chain termination method (Sanger *et al.*, 1977).

Acknowledgements

We thank Karin Harrison for excellent technical assistance, Dr. Barbara van Leeuwen for valuable discussions and criticism of the manuscript, and Drs. Ronne, Bradshaw, Fahnestock and Shooter for communication of data prior to publication. B.A.E. is the recipient of a National Health and Medical Research Council Postdoctoral Fellowship.

References

Benoist,C., O'Hare,K., Breathnach,R. and Chambon,P. (1980) *Nucleic Acids Res.*, **8**, 127-142.
Benton,W.D. and Davis,R.W. (1977) *Science (Wash.)*, **196**, 180-182.
Berger,E.A. and Shooter,E.M. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 3647-3651.
Bode,W. (1979) *J. Mol. Biol.*, **127**, 357-374.
Bode,W. (1983) *Naturwissenschaften*, **66**, 251-825.
Bradshaw,R.A. (1978) *Annu. Rev. Biochem.*, **47**, 191-216.
Corden,J., Wasylyk,B., Buchwalder,A., Sassone-Corsi,P., Kedinger,C. and Chambon,P. (1980) *Science (Wash.)*, **209**, 1406-1414.
Dayhoff,M.O. (1978) *Atlas of Protein Sequence and Structure*, Vol. **5**, Suppl. 3, published by Georgetown University Press, Washington, DC.
Enquist,L. and Sternberg,N. (1979) *Methods Enzymol.*, **68**, 281-298.
Frischauf,A.-M., Lehrach,H., Poustka,A. and Murray,N. (1983) *J. Mol. Biol.*, **170**, 827-842.
Greene,L.A., Shooter,E.M. and Varon,S. (1968) *Proc. Natl. Acad. Sci. USA*, **60**, 1383-1388.
Greene,L.A., Shooter,E.M. and Varon,S. (1969) *Biochemistry (Wash.)*, **8**, 3735-3741.
Howles,D.N., Dickinson,D.P., DiCaprio,L.L., Woodworth-Gutai,M. and Gross,K.W. (1984) *Nucleic Acids Res.*, **12**, 2791-2805.
Isackson,P.J. and Bradshaw,R.A. (1984) *J. Biol. Chem.*, **259**, 5380-5383.
Isackson,P.J., Ullrich,A. and Bradshaw,R.A. (1985) *Biochemistry (Wash.)*, in press.
Karin,M., Cathala,G. and Nguyen-Huu,M.C. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 4040-4044.
Karin,M., Haslinger,A., Holtgreve,H., Richards,R.I., Krauter,P., Westphal, H.M. and Beato,M. (1984) *Nature*, **308**, 513-519.
Krieger,M., Kay,L.M. and Stroud,R.M. (1974) *J. Mol. Biol.*, **83**, 209-230.
Mason,A.J., Evans,B.A., Cox,D.R., Shine,J. and Richards,R.I. (1983) *Nature*, **303**, 300-307.
Messing,J. (1983) *Methods Enzymol.*, **101**, 20-78.
Poe,M., Wu,J.K., Florance,J.R., Rodkey,J.A., Bennett,C.D. and Hoogsteen,K. (1983) *J. Biol. Chem.*, **258**, 2209-2216.
Proudfoot,N.J. and Brownlee,G.G. (1976) *Nature*, **263**, 211-214.
Richards,R.I., Catanzaro,D.F., Mason,A.J., Morris,B.J., Baxter,J.D. and Shine,J. (1982) *J. Biol. Chem.*, **257**, 2758-2761.
Richards,R.I., Heguy,A. and Karin,M. (1984) *Cell*, **37**, 263-272.
Ronne,H., Lundgren,S., Severinsson,L., Rask,L. and Peterson,P.A. (1983) *EMBO J.*, **2**, 1561-1564.
Ronne,H., Anundi,H., Rask,L. and Peterson,P.A. (1984) *Biochemistry (Wash.)*, **23**, 1229-1234.
Sanger,F., Nicklen,S. and Coulson,A.R. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 5463-5467.
Scott,J., Selby,M., Urdea,M., Quiroga,M., Ball,G.I. and Rutter,W.J. (1983) *Nature*, **302**, 538-540.
Silverman,R.E. and Bradshaw,R.A. (1982) *J. Neurosci. Res.*, **8**, 127-136.
Southern,E.M. (1975) *J. Mol. Biol.*, **98**, 503-517.
Stach,R.W., Pignatti,P.-F., Baker,M.E. and Shooter,E.M. (1980) *J. Neurochem.*, **34**, 859-865.
Swift,G.H., Dagorn,J.-C., Ashley,P.L., Cummings,S.W. and MacDonald, R.J. (1982) *Proc. Natl. Acad. Sci. USA*, **79**, 7263-7267.
Tschesche,H., Mair,G., Godec,G., Fiedler,F., Ehret,W., Hirschauer,C., Lemon,M. and Fritz,H. (1979) *Adv. Exp. Med. Biol.*, **120**, 245-260.
Thomas,K.A., Baglan,N.C. and Bradshaw,R.A. (1981) *J. Biol. Chem.*, **256**, 9156-9166.
Ullrich,A., Gray,A., Berman,C. and Dull,T.J. (1983) *Nature*, **303**, 821-825.
Varon,S., Nomura,J. and Shooter,E.M. (1968) *Biochemistry (Wash.)*, **7**, 1296-1303.
Walker,M.D., Edlund,T., Boulet,A.M. and Rutter,W.J. (1983) *Nature*, **306**, 557-561.
Yanker,B.A. and Shooter,E.M. (1982) *Annu. Rev. Biochem.*, **51**, 849-868.

Received on 17 October 1984