Biochimica et Biophysica Acta 909 (1987) 10–20
Elsevier

BBA 91710

The structure of a human neurofilament gene (NF-L): a unique exon-intron organization in the intermediate filament gene family

Jean-Pierre Julien, Frank Grosveld, Karina Yazdanbaksh, David Flavell, Dies Meijer and Walter Mushynski *

Laboratory of Gene Structure and Expression, National Institute for Medical Research, Mill Hill, London (U.K.)

(Received 7 November 1986)

Key words: Neurofilament gene; Nucleotide sequence; Exon-intron organization; Cloning; Cosmid library; (Human DNA)

We have cloned and determined the nucleotide sequence of the human gene for the neurofilament subunit NF-L. The cloned DNA contains the entire transcriptional unit and generates two mRNAs of approx. 2.6 and 4.3 kb after transfection into mouse L-cells. The NF-L gene has an unexpected intron-exon organization in that it entirely lacks introns at positions found in other members of the intermediate filament gene family. It contains only three introns that do not define protein domains. We discuss possible evolutionary schemes that could explain these results.

Introduction

Cytoplasmic intermediate filaments have been divided into five subclasses based on their biochemical properties, immunological specificity and tissue distribution: keratin filaments in epithelial cells, vimentin filaments in cells of mesenchymal origin, desmin in muscle cells, glial filaments in astrocytes and neurofilaments in neurons (for review see Refs. 1, 2). The different types of intermediate filament protein share common structural features; sequence data indicate that they contain a homologous α-helical domain of conserved length capable of forming coiled-coils, which is flanked by amino- and carboxy-terminal domains of variable size and composition [3–10].

Mammalian neurofilaments are composed of three neuron-specific proteins with apparent molecular weights of 68 000 (NF-L), 145 000 (NF-M) and 200 000 (NF-H) on SDS-gel electrophoresis [11,12]. Neurofilament proteins share homologous α-helical domains with other types of intermediate filament protein, but contain a long extension at their COOH-terminus which varies in size between the three subunits [13,14]. In each of the neurofilament proteins this so-called tail domain contains high levels of charged amino-acid residues [4,13,15,16] as well as multiple phosphorylation sites in NF-M and NF-H [17–20].

Non-neuronal intermediate filament genes that have been analyzed to date reveal a remarkable conservation of intron positions. The vimentin, desmin and glial fibrillary acidic protein genes each contains eight introns at identical positions, six of the introns being located within the regions encoding α-helical sequences [9,10,21]. A majority of the introns in the less closely related keratin genes occur at similar or identical positions [22–27].

It was therefore concluded that introns must have been present in an ancestral intermediate

* Present address: Department of Biochemistry, McGill University, Montreal, Canada.

Correspondence: (present address): J-P. Julien, Institut du Cancer de Montréal, Hôpital Notre-Dame, 1560 est, Sherbrooke, Montréal, Canada, H2L 4M1.

filament gene prior to duplication and subsequent divergence of the intermediate filament gene family. The preferential conservation of introns within the α-helical domain reflects the evolutionary stability of this region [25].

We report here the isolation of the genomic copy of the human NF-L gene using an NF-L cDNA probe. After transfection into mouse L-cells the cloned gene was appropriately transcribed into two mRNAs. The human NF-L gene was fully sequenced and its structure determined. The gene contains only three introns which are located at sites found in the corresponding murine gene [28]. To explain the unique exon-intron organization of the NF-L gene, Lewis and Cowan proposed a mechanism by which the reverse transcription of an ancestral intermediate filament mRNA was followed by integration into the genome [28]. An equally plausible explanation which is also discussed here proposes that the primordial neurofilament gene diverged before the rest of the intermediate filament gene family.

Materials and Methods

Isolation and sequencing of the human gene encoding the NF-L protein

Screening of a cosmid library constructed from a partial *Mbo*I digest of human DNA ligated to the vector pTCF [29] was carried out using a 580 bp *Bgl*II/*Xho*I fragment of rat NF-L cDNA [15]. The probe was ^{32}P-labeled by means of the Klenow fragment used in conjunction with hexanucleotide primers [30]. One positive clone, designated pHNFL, was isolated and mapped by standard double restriction enzyme digestion.

Sequencing of the NF-L gene was initially carried out by the M13-dideoxy chain termination procedure [31] in conjunction with the shot-gun cloning of 300–500 bp DNA fragments that were generated by sonication of the circularized 6.5 kb *Eco*RI fragment of pHNFL.

DNA transfection

The cosmid DNA containing the NF-L gene (pHNFL) was introduced into mouse L-cells by calcium phosphate coprecipitation [32]. Following transfection with pHNFL, the cells were harvested after 2 days in culture and analyzed for the presence of human NF-L mRNA by Northern blot analysis.

DNA and RNA blot analysis

Human DNA was digested with various restriction endonucleases and fractionated on 0.7% agarose gels. The DNA was transferred to nitrocellulose [33] and the blots were hybridized as described previously [15]. Isolation of total RNA from cells in culture and from the mouse brain was carried out by the guanidinium/CsCl method [34]. RNA samples were fractionated by electrophoresis on 1.0% agarose gels in the presence of formaldehyde, blotted and hybridized as described previously [15].

Results

Isolation and sequencing of the human NF-L gene

To isolate the human NF-L gene we used a

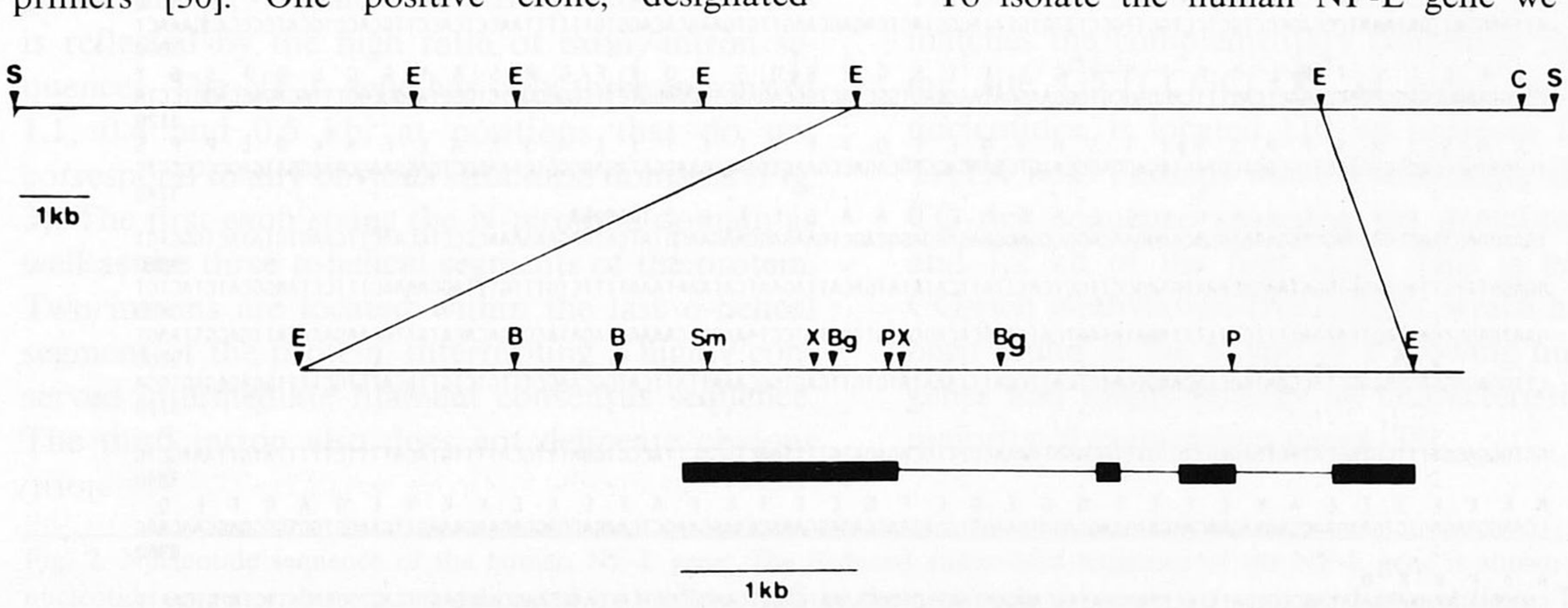

Fig. 1. Restriction cleavage map and intron/exon organization of the human NF-L gene. Abbreviations of restriction enzyme names are: B, *Bam*HI; Bg, *Bgl*II; C, *Cla*I; E, *Eco*RI; P, *Pvu*II; Sm, *Sma*I; S, *Sal*I; X, *Xho*I. The lowest line is a schematic representation of the NF-L gene. Exons are depicted as boxes, introns as lines.

```
GCAGCTCCTCGGGCCGTAGCTCGACCCCGCCTTCCCTTTTCCGCAGAATCCTCGCCTTGGCTGCAGCAGCGCGCTGCCCCCACTGGCCGGCGTGCCGTGATCGATCGCAGGCTGCGTCAG
                                                                                                                        120
                                  CAP SITE
GACCTCCCGGCGTATAAATAGGGGTGGCAGAACGGCGCCGAGCCGCACACAGCCATCCATCCTCCCCCTTCCCTCTCTCCCCTGTCCTCTCTCTCCGGGCTCCCACCGCCGCCGGGGAGC
                                                                                                                        240
                  S  S  F  S  Y  E  P  Y  Y  S  T  S  Y  K  R  R  Y  V  E  T  P  R  V  H  I  S  V  R  S  G  Y  S  T  A
ACCGGCCGCCAACCAATGAGTTCCTTCAGCTACGAGCCGTACTACTCGACCTCCTACAAGCGGCGCTACGTGGAGACGCCCCGGGTGCATATCAGCGTGCGCAGCGGCTACAGCACCGCA
                                                                                                                        360
  R  S  A  Y  S  S  Y  S  A  P  V  S  S  S  L  S  V  R  R  S  Y  S  S  S  S  G  S  L  M  P  S  L  E  N  L  D  L  S  Q  V
CGCTCAGCTTACTCAAGCTACTCGGCGCCGGTGTCTTCCTCGCTGTCCGTGCGCCGCAGCTACTCCTCCAGCTCTGGATCGTTGATGCCCAGTCTGGAGAACCTCGACCTGAGCCAGGTA
                                                                                                                        480
  A  A  I  S  N  D  L  K  S  I  R  T  Q  E  K  A  Q  L  Q  D  L  N  D  R  F  A  S  F  I  E  R  V  H  E  L  E  Q  Q  N  K
GCCGCCATCAGCAACGACCTCAAGTCCATCCGCACGCAGGAGAAGGCGCAGCTCCAGGACCTCAATGACCGCTTCGCCAGCTTCATCGAGCGCGTGCACGAGCTGGAGCAGCAGAACAAG
                                                                                                                        600
  V  L  E  A  E  L  L  V  L  R  Q  K  H  S  E  P  S  R  F  R  A  L  Y  E  Q  E  I  R  D  L  R  L  A  A  E  D  A  T  T  N
GTCCTGGAAGCCGAGCTGCTGGTGCTGCGCCAGAAGCACTCCGAGCCATCCCGCTTCCGGGCGCTGTACGAGCAGGAGATCCGCGACCTGCGCCTAGCGGCGGAAGATGCCACCACCAAC
                                                                                                                        720
  E  K  Q  A  L  R  G  E  R  E  E  G  L  E  E  T  L  R  N  L  Q  A  R  Y  E  E  E  V  L  S  R  E  D  A  E  G  R  L  M  E
GAGAAGCAAGCGCTCCGAGGCGAGCGCGAAGAAGGGCTGGAGGAGACCCTGCGCAACCTGCAGGCGCGCTATGAAGAGGAGGTGCTGAGCCGCGAGGACGCCGAGGGCCGGCTGATGGAA
                                                                                                                        840
  R  R  K  G  A  D  E  A  A  L  A  R  A  E  L  E  K  R  I  D  S  L  M  D  E  I  S  F  L  K  K  V  H  E  E  E  I  A  E  L
CGCCGCAAAGGCGCCGACGAGGCGGCGCTCGCTCGCGCCGAGCTCGAGAAGCGCATCGACAGCTTGATGGACGAAATCTCTTTTCTGAAGAAAGTGCACGAAGAGGAGATCGCCGAACTG
                                                                                                                        960
  Q  A  Q  I  Q  Y  A  Q  I  S  V  E  M  D  V  T  K  P  D  L  S  A  A  L  K  D  I  R  A  Q  Y  E  K  L  A  A  K  N  M  Q
CAGGCGCAGATCCAGTACGCGCAGATCTCCGTGGAGATGGACGTGACCAAGCCCGACCTTTCCGCCGCGCTCAAGGACATCCGCGCGCAGTACGAGAAGCTGGCCGCCAAGAACATGCAG
                                                                                                                        1080
  N  A  E  E  W  F  K  S  R  F  T  V  L  T  E  S  A  A  K  N  T  D  A  V  R  A  A  K  D  E  V  S  E  S  R  R  L  L  K  A
AACGCTGAGGAATGGTTCAAGAGCCGCTTCACGGTGCTGACCGAGAGCGCCGCCAAGAACACCGACGCCGTGCGCGCCGCCAAGGACGAGGTGTCGGAGAGCCGTCGTCTGCTCAAGGCC
                                                                                                                        1200
  K  T  L  E  I  E  A  C  R  G  M  N  E  A  L  E  K  Q  L  Q  E  L  E  D  K  Q  N  A  D  I  S  A  M  Q  INTRON1
AAGACCCTGGAAATCGAAGCATGCCGGGGCATGAATGAAGCGCTGGAGAAGCAGCTGCAGGAGCTGGAGGACAAGCAGAACGCCGACATCAGCGCTATGCAGgtgcggcacggccagaaa
                                                                                                                        1320
CACAGGGGGGCGGGGAACTCGAGCAAGGGGGGGAGTTGGTGCGCCCAGAAAGCGAAACCAGGGGTGGTGCGGCTGCCCAGCTCTTAGGGATAGGGCTTGGCTCCTTGGCCACTGTGTGGA
                                                                                                                        1440
GGGGTGGGGCTCTTGAGGGGCGTGAGTGCGGGCGCCACTGTAGTCCGGGAGTGACTGCTCCGCGTGCTGCACCGGCGTTCCGCATTAAAGCTGCCCGACCCTTGTTGGGTGGGGAGGGGA
                                                                                                                        1560
AGACGTGGGAATTGGGCGTTGCCTCCGACTGCAGTGAGATCAGCTCTCTACTGACCTGCGTTGACCACGAGTTACTTTGCAGCGACTATCGGATCGTCTAGTTAATAAATAGTACGAGTG
                                                                                                                        1680
AACTAACTCTCAATTAATTCTGAAGGATTACTGTGACCAGCATGCTTTATGACTAGTTTTACCAACCACCTCCCTTCCTTTATTTAGTAGGTAGACAGGAAAATAGTCAACATTGTTTTA
                                                                                                                        1800
GGTAGTTAACTAGTGATGTTCATAGTAAACCATTTCCTTTTACCTTTTTTTTTTCTTTTTTTCTTTATGTGTAAAATCTTCTACAACATTTCTGTTTAAACATCTCCATCTTCTGGGGAG
                                                                                                                        1920
TAGAAAAAATACAATTTTAAAAAGATCTCCATTTTAAAACATCTCCGTCTTCTGGGGAGTAGAAATTTTTTTCTTCTTCTGGGGAGTAGAAAAAATAATTTAGATACATAGGAAATATTT
                                                                                                                        2040
CATAGAAAATAATTTTTTTCTTTTTTTTGTTTACATCTGGTATTTTCTTCTCATAAAGAAAGGCATTAGTTTCCTGGCATGTAACCCAGCTAAAGAAGAGTAATCAGTGAATGAGAGACA
                                                                                                                        2160
CAGTTTTTCTATCAACTTAGTCTGTTTTTCTATCAACTTAGTCTGTTTGCATGCATTTTATGATGATCATTAAACAGTATTAAGTAAAGAAACAGAAGAACAGAATTTTCGTCCATCTTT
                                                                                                                        2280
TTTTTCATCTCAGGCTTCATGAACTTGGGTATTTTAGGCATGAAGGTTTTTCAAAAGATACAGGAAGTTATTCTAGGAGAGATTTTATCAAAGTGTGCACCTTGATTTTAATCGAAACTA
                                                                                                                        2400
                                                                                                  D  T  I  N  K  L  E  N  E  L  R  T
GGCCTTTGCAACTACACTACAGTAAAATAATAGAAGGGATTTATGCTCGGATTTTTTTTTGTTTTATTTTGTCTTCAAACAGgacacgatcaacaaattagaaaatgaattgaggacc
                                                                                                                        2520
  T  K  S  E  N  A  R  Y  L  K  E  Y  Q  D  L  L  N  V  K  M  A  L  D  I  E  I  A  A  Y  R  INTRON2
ACAAAGAGTGAAATGGCACGATACCTAAAAGAATACCAAGACCTCCTCAACGTGAAGATGGCTTTGGATATTGAGATTGCTGCTTACAGgtgaaaatagagggcaaagacagcagccat
                                                                                                                        2640
TAAACCTTAGGAAGAAAATCAGATCCCATTTAAAGTTATGTTGGATCAGAAACCTTCAATAATAGTCCTTTTGAAATAATGAAGTGTTAGTTTTTGGCTTCTTCCAAGAAGAGGGTATTT
                                                                                                                        2760
AGATATATAAGAATTTAACCCTGTAATTAGAGTCCTGTTTTTATCTTGTCATTACACTTTAAATCTAATAGGAGTGATTTATTTATATTTTTTCTGGTCTCCATCAAAAGATCCCCAGGC
                                                                                                                        2880
                                                                                                                        K  L
ATTAAGTATTGATAAATCCCAGCCCTGCTCCTGCTTGCCTTTGTGTTTAGGGTACTCAGAGCAAGTTGTGAAACACAGGTGTTTTTTAACCTCACCTTGCACCTGCATCCCCAGgaaact
                                                                                                                        3000
  L  E  G  E  E  T  R  L  S  F  T  S  V  G  S  I  T  S  G  Y  S  Q  S  S  Q  V  F  G  R  S  A  Y  G  G  L  Q  T  S  S  Y
CTTGGAAGGCGAGGAGACCCGACTCAGTTTCACCAGCGTGGGAAGCATAACCAGTGGCTACTCCCAGAGCTCCCAGGTCTTTGGCCGATCTGCCTACGGCGGTTTACAGACCAGCTCCTA
                                                                                                                        3120
  L  M  S  T  R  S  F  P  S  Y  Y  T  S  H  V  Q  E  E  Q  T  E  V  E  E  T  I  E  A  S  K  A  E  E  A  K  D  E  P  P  S
TCTGATGTCCACCCGCTCCTTCCCGTCCTACTACACCAGCCATGTCCAAGAGGAGCAGACCGAAGTGGAGGAAACCATTGAGGCGTCTAAGGCTGAGGAAGCCAAGGATGAGCCCCCCTC
                                                                                                                        3240
  E  G  E  A  E  E  E  E  K  D  K  E  E  A  E  E  E  E  A  A  E  E  E  E  INTRON3
TGAAGGAGAAGCCGAGGAGGAGGAGAAGGACAAGGAAGAGGCCGAGGAAGAGGAGGCAGCTGAAGAGGAAGAAGgtatgataagaaaaaaccccctgcaacttcaagtgtaaactgggtgt
                                                                                                                        3360
GGAGATTTGTTAGGAGGTGGATAAGACAAATGAAGCCTTGCTCATTTATTCATATATGACATTAGAATCATAAATAAATTTTCTGTTTGTTTAGCAAAACTTTCCTAAGGCATCTACTCT
                                                                                                                        3480
GAATGAGGTGATTGGTCAAAATTTTCATTTTTTAATATAATCATTTAACACAGCAGGTTGGTGTCCTAAAGAACAAAAATAGATACCAGACACATAATGAAAGAAATATTGAGGTTAAGT
                                                                                                                        3600
CTTGGAGAGGAGCAGAGCTTCCCATACCTAGAAGTGATCTCATTCGATTTAAATATGTGTTCAGTGGCAAATTATTCATGGCAAGCTTTGTCTGTTACATGTGCTTTTGGAGAGAGTGGA
                                                                                                                        3720
                                                                                                                        A
GCTGGGAGGTTTTGGTAGCATTCTGACAGTTGTGTTTGCAAATAAAACCTTTGCAGACATGTTTTGACTGGACTTACCCTGGATTTGCATTTTGTACATTTTCTTTTTATGTTAAAGctg
                                                                                                                        3840
  A  K  E  E  S  E  E  A  K  E  E  E  E  G  G  E  G  E  E  G  E  E  T  K  E  A  E  E  E  E  K  K  V  E  G  A  G  E  E  Q
CCAAGGAAGAGTCTGAAGAAGCAAAAGAAGAAGAAGAAGGAGGTGAAGGTGAAGAAGGAGAGGAAACCAAAGAAGCTGAAGAGGAGGAGAAGAAAGTTGAAGGTGCTGGGGAGGAACAAG
                                                                                                                        3960
  A  A  K  K  K  D
CAGCTAAGAAGAAAGATTGAACCCCCATTTCCTTAATTATTTCAGGAATAATTCTCCCGAAATCAGGTCAACCCCATCACCAACCAACCAACCAGTTGAGTTCCAGATTCTATGTGAATT
                                                                                                                        4080
AAAAAGTCAATATATGTATAATTCTGAGATGACTTAGGTTGGACATTCAATGTTGTGCTATGAATTTCCTCTTTATGCAGAGTATCTGTTTGCTTGCAGAGTGGCTTTCGGCTTGCTGCC
                                                                                                                        4200
AGCCTGTGCATGGTCCACGCTTATGAGTTCAGGATCTACGGCAATGTGAATCATTCAGATGTTTACAATAAAAAACACCACATGAGTAAATGAATTCACTAATGTTAATGTTAAACTTCA
                                                                                                                        4320
TGGAAAAGTAGTCCTTTGAACCTTCGGTGGTTAGCAATTAAAGACCCTGAGTTATGTGCAATAAATAGTAAATAAAGTTATACCGAATGATGTATTTTTTGCCGTGGTTGTTACCTAATT
                                                                                                                        4440
AAAATACCTTAAAGATGGCACCAATATAAAGTGTGTGCCAGTGAACTATTGACCTCCAATTTTTTAAAAAGCCGAAATTTTAACAATTACCAATACTTTTTT
                                                                                                                        4542
```

*Bgl*II/*Xho*I cDNA probe of rat NF-L [15] to screen a cosmid library constructed from a partial MboI digest of human DNA ligated to the vector pTCF [29]. Approximately $7 \cdot 10^5$ recombinants were screened and one clone was isolated that hybridized strongly to the cDNA probe.. This clone, designated pHNFL, contains a 34 kb insert that generates all of the NF-L hybridization bands in total human DNA (not shown).

A 21.5 kb *Sal*I fragment from the genomic insert was partially mapped, as shown in Fig. 1. Hybridization with *Eco*RI/*Bgl*II fragments corresponding to the 5′ and 3′ segments of the mouse NF-L cDNA [35] suggested that all or most of the human NF-L gene was located on the single 6.5 kb *Eco*RI fragment shown in Fig. 1, which corresponds to the *Eco*RI hybridization band in total human DNA (data not shown). A detailed restriction map of this *Eco*RI fragments is also presented in Fig. 1.

To elucidate the structure and the coding information of the human NF-L gene, the 6.5 kb *Eco*RI fragment was sequenced by the dideoxy chain termination method (Fig. 2). The coding sequences of the NF-L gene are highly conserved, thereby allowing us to deduce the exon-intron organization of the human gene by comparing the sequence with the corresponding full length cDNA sequence of mouse NF-L [36]. The 541 amino-acid residues predict a molecular weight of approx. 61 000 for the human NF-L protein.

In comparison to other intermediate filament genes, the NF-L gene is relatively small, and this is reflected by the high ratio of exon/intron sequences. The NF-L gene has only three introns of 1.1, 0.4 and 0.5 kb, at positions that do not correspond to any obvious structural domains (Fig. 3). The first exon spans the N-terminal domain as well as the three α-helical segments of the protein. Two introns are located within the last α-helical segment of the protein, interrupting a highly conserved intermediate filament consensus sequence. The third intron also does not delineate obvious

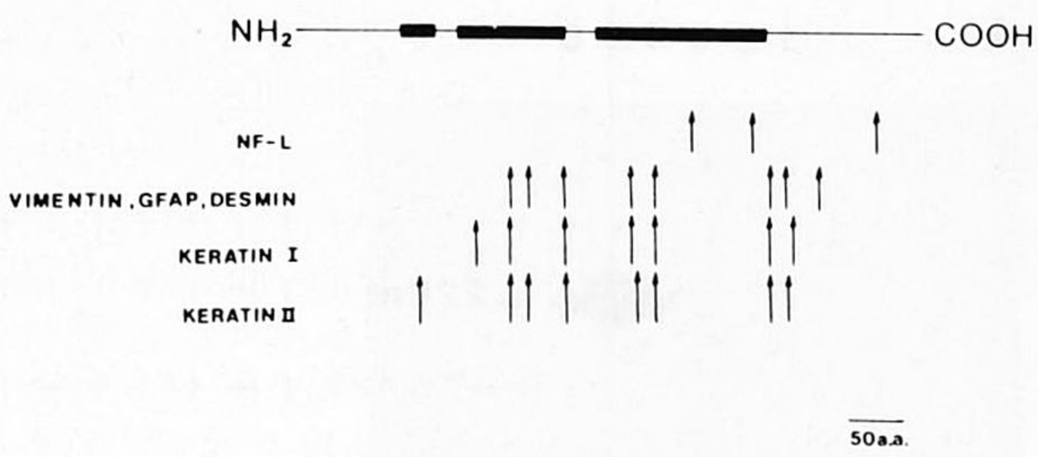

Fig. 3. Relationship between the intron positions of intermediate filament genes and the major structural regions of the proteins. Like all intermediate filament proteins, the NF-L protein is composed of a central α-helical domain flanked by non-α-helical NH_2- and COOH-terminal domains. The α-helical regions are boxed. Arrows indicate the intron positions of each gene within the corresponding sequence of the NF-L protein.

subdomains of the carboxy terminal domain of the protein.

S1 nuclease protection experiments indicate that the cap site is located 100 nucleotides upstream of the ATG initiation codon (Fig. 4A). The human NF-L gene promoter contains a typical TATAA sequence 30 nucleotides upstream from the cap site. Other upstream elements such as the CAAT box or Sp1 transcription fractor binding sites [37] can be found in sequences which share only a partial homology with the consensus sequence. The sequence GATCGATC, which is homologous to the complementary CAAT box consensus sequence GATTGACC matching six out of eight base-pairs, is located 35 bp upstream from the TATA box. The sequence ACCCCGCCTT, which matches the complementary consensus Sp1 binding site GTTCCGCCCC in eight out of ten nucleotides, is located 110 bp uptream from the TATA box. Perhaps more interestingly there is a CG-rich sequence spanning the promoter region and 1.2 kb of the first exon. This is typical of CG-rich methylation-free islands, which have now been found at the 5′ end of a growing number of genes and might actually be characteristic of the majority of mammalian genes [38].

Fig. 2. Nucleotide sequence of the human NF-L gene. The deduced amino-acid sequence of the NF-L gene is shown above the nucleotide sequence. Introns are delineated by brackets. 97% of the nucleotide sequence was sequenced several times. The arrows indicate the intron positions in the corresponding sequence of vimentin [9] desmin [10] and glial fibrillary acidic protein [21] genes. The Sp1 binding site ACCCCGCCTT and the sequence GATCGATC which is homologous to complementary CAAT box sequence GATTGACC are located 110 bp and 35 bp upstream from the TATAA box, respectively.

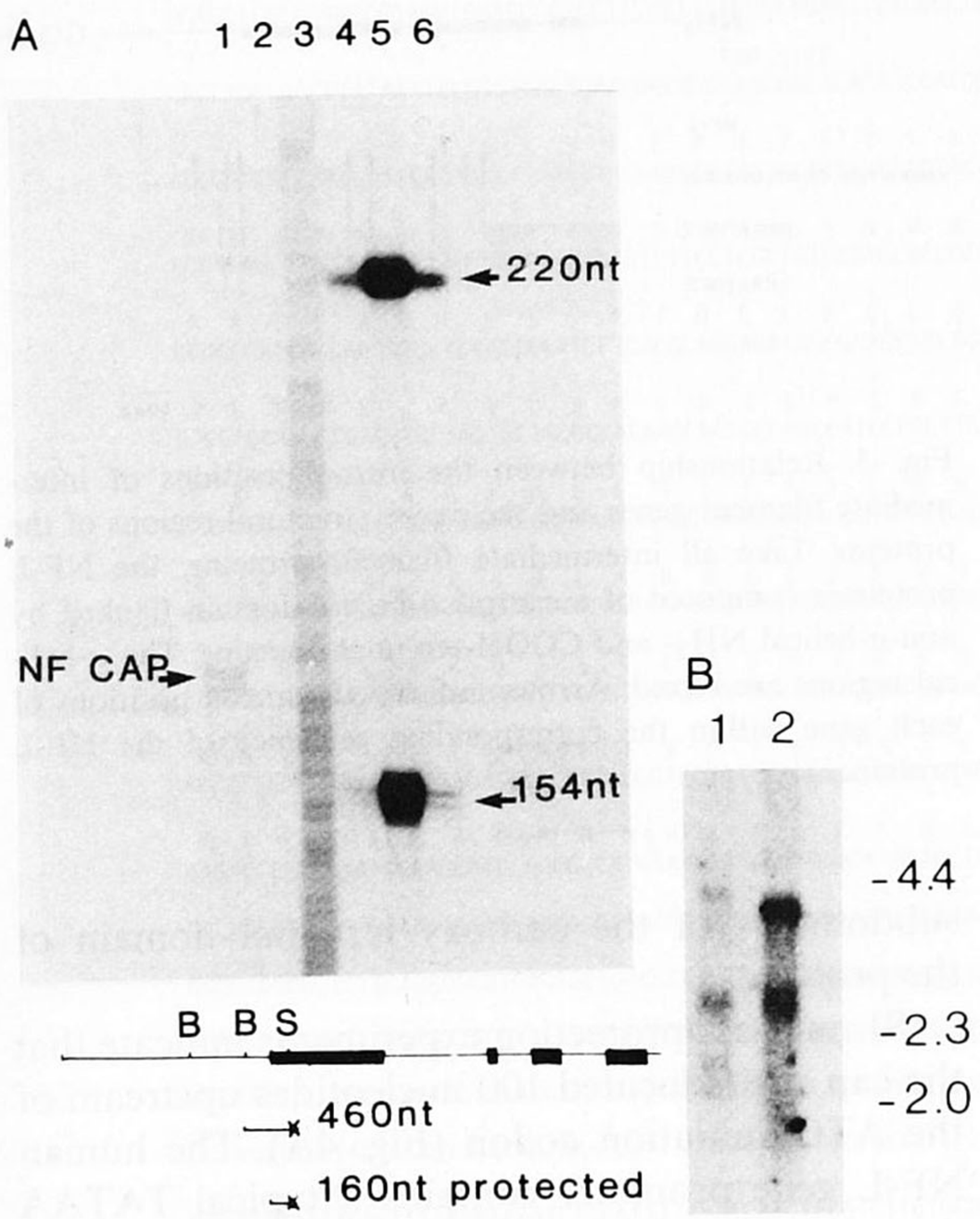

Fig. 4. (A) Mapping the site of transcription initiation by S1 nuclease analysis. Total RNA from human brain (35 μg, lane 1) and HeLa cells (35 μg, lane 2) was annealed to a 5′-end-labelled probe and subjected to S1 nuclease analysis [65]. The identical fragment was subjected to G and A degradation reactions [66] (lane 3). Lanes 4-6 pBR322 × *Hin*fI marker fragments. The diagram shows the *Sma*I end-labelled probe used for S1 nuclease analysis. The exons of the NF-L gene are represented by filled-in boxes enzymes used: B, *Bam*HI; S, *Sma*I. (B) Blot hybridization of RNA from mouse L-cells transfected with DNA containing the human gene for NF-L. Total RNA was extracted from mouse L-cells two days after transfection with DNA from pHNFL. About 5 μg of total RNA was resolved on a denaturing 1% agarose/formaldehyde gel (lane 1) alongside a sample containing 5 μg of total mouse brain RNA (lane 2). Size markers in kb are shown on the right.

Chromosomal localization of the human NF-L gene

A 4.3 kb 5′ *Xma*I-*Eco*RI-3′ probe corresponding to the NF-L gene (see below) was used as a probe in Southern blot hybridizations to DNA isolated from a human-hamster hybrid cell panel [39]. A positive 6.5 kb *Eco*RI hybridization signal was obtained whenever human chromosome 8 was present in the hybrid cells (Table I). In addition, all the cells showed a 3.6 kb *Eco*RI signal for the hamster neurofilament gene. We conclude that a single human NF-L gene is located on chromosome 8.

Expression of the human NF-L gene in mouse L-cells

The single copy NF-L gene has been shown to produce two mRNA species of approx. 2.5 and 3.5 kb in both mouse [28] and rat brain [15]. This situation is analogous to that for the single copy chicken vimentin gene which gives rise to two mRNA species because of occasional read through the first of two sets of tandem polyadenylation sites [40,41].

Our initial assessment of whether the cosmid pHNFL contained the entire functional NF-L gene was carried out by DNA-mediated gene transfer. Mouse L-cells were transfected with the cosmid pHNFL and RNA extracted 2 days later was subjected to Northern blot analysis. Fig. 4B shows that transfected mouse L-cells, which do not express the murine NF-L gene, contained two human NF-L transcripts of about 2.6 and 4.3 kb, indicating that the pHNFL clone contains the entire transcriptional unit of the human NF-L gene. The small mRNA species are of identical size in both human and mouse, while the other transcript is larger in humans than in mice. Northern blot analysis of a small amount of human brain RNA indeed showed two human NF-L mRNA species of 2.6 and 4.3 kb, despite the obvious difficulty in obtaining high quality human brain mRNA (not shown). These results indicate that the introduced human NF-L gene is transcribed from its own promoter and processed correctly in L-cells.

Discussion

We report here the cloning of the human gene encoding the NF-L protein. The dot-matrix comparison in Fig. 5A reveals extensive sequence homology between the human and the recently described mouse NF-L genes [28]. In both species the exons and the 5′ untranslated sequences are highly conserved (90% homology). The sizes and positions of introns have been maintained but the intron sequences evolved with considerable drift. Fig. 5B shows a dot-matrix comparison between the nucleotide sequences of the human NF-L and NF-M (Myers, M.W., Lazzarini, R.A., Lee, V.M.-

TABLE I

HUMAN CHROMOSOME CONTENT OF HUMAN CELL HYBRID AS DETERMINED BY KARYOTYPE ANALYSIS

Columns 1–Y: + and − indicate the presence or absence of chromosomes 1 to Y; ± indicates the presence of a chromosome but in 10% of the hybrid cells. The right-hand column shows the human NF-L scoring of the cell lines by Southern blot hybridization.

Hybrid	Chromosome complement																							
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	X	Y	Presence of NF-L
E36 4.1.5	+	+	+	±	+	+	±	+	+	+	+	+	±	+	−	+	+	+	+	+	−	+	±	−	+
E36 2.9.5	+	+	+	+	−	+	−	±	−	+	−	+	−	+	+	+	−	+	+	−	+	+	+	−	+
17CB3B	+	+	−	+	+	+	+	−	+	−	−	+	+	+	+	+	+	−	+	+	+	−	−	−	−
17 CB2C	−	−	−	−	−	+	+	−	−	−	−	−	−	−	−	−	+	−	−	−	+	−	−	−	−
17CB22C	−	±	+	±	−	+	+	+	+	+	−	−	+	−	−	+	±	−	−	+	+	+	+	−	+
16CB17	+	−	+	+	−	−	−	+	−	−	−	+	−	+	−	+	+	+	+	−	+	+	−	−	+
16CB18	−	±	+	+	−	−	±	−	+	−	+	−	−	+	+	+	−	−	−	+	±	+	+	−	−
17CB17A	−	−	−	±	−	−	−	−	−	−	+	−	−	−	−	+	±	−	+	−	−	−	−	−	−
17CN17B	−	−	+	+	±	+	−	+	+	+	+	+	−	−	−	−	+	+	+	−	+	+	−	−	+
17CB16C	+	−	+	+	−	+	−	+	+	−	−	+	−	±	−	−	+	+	±	+	−	−	−	−	+
17CB10	−	−	−	+	−	+	−	−	+	−	+	+	−	+	−	−	+	+	±	+	+	−	±	±	−
16CB15A	+	−	+	+	−	+	-	−	−	−	+	−	+	+	−	−	+	−	+	+	+	+	−	−	−
17CB29B	−	−	−	−	−	+	+	+	−	+	+	+	+	−	±	−	+	−	+	−	+	−	−	−	+
16CB16A	−	−	−	+	−	+	−	+	+	−	−	−	−	−	−	+	+	-	−	−	+	+	−	−	+
16CB1	+	−	+	+	+	+	−	−	−	−	+	−	−	+	+	+	+	+	+	+	−	+	−	−	−
17CB8B	−	−	−	+	+	+	−	+	+	+	+	+	±	−	−	+	+	−	+	±	+	+	±	−	+
17CB21C	−	−	−	+	−	−	+	−	+	−	−	+	−	+	−	−	+	−	+	+	+	−	±	−	−
17CB15B	−	±	+	+	+	−	−	−	−	+	+	+	+	−	−	+	+	±	±	+	+	+	−	+	−

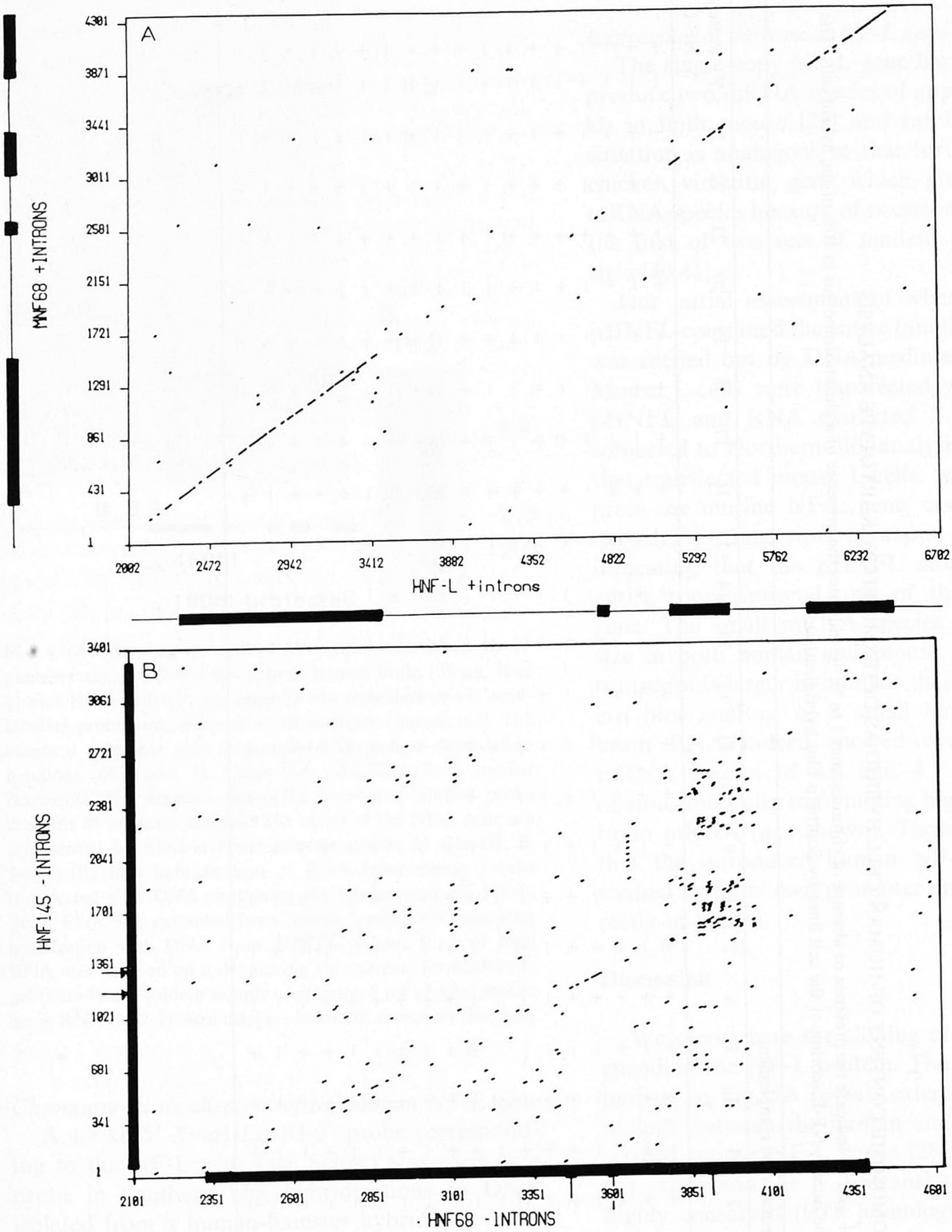

Fig. 5. (A) Dot-matrix comparison between the nucleic acid sequences of the human and the mouse NF-L gene [28]. Schematic representations of the human and mouse NF-L genes are shown on the horizontal and vertical axes, respectively. Exons are depicted as boxes; introns as lines. Nucleotide 2170 of the human NF-L corresponds to the first nucleotide in Fig. 2. (B) Dot-matrix comparison between the human NF-L and NF-M genes (Meyers, M.W., Lazzarini, R.A., Lee, V.M.-Y., Schlaepfer, W.W. and Nelson, D.L., unpublished data). The NF-L sequences are on the horizontal axis. The analysis does not include intron sequences.

Y., Schlaepfer, W.W. and Nelson, D.L., unpublished data). The analysis does not include the intron sequences. The highest degree of homology is observed at the two ends of the α-helical-rich regions (nucleotides 2700–2850 and 3500–3600 in the NF-L gene), which represent consensus sequences among intermediate filament proteins [13]. The two neurofilament genes contain in their corresponding carboxy-terminal domains GAG repeats that encode glutamic acid. This is illustrated by the clustering of dots between nucleotides 3600–4000 in the NF-L gene (Fig. 5B).

It is rather surprising that the NF-L gene completely lacks the introns present in other types of the intermediate filament gene. This anomalous exon/intron pattern appears to occur also in other members of the neurofilament gene family. A preliminary analysis of the NF-H gene shows at least two intron positions in the conserved α-helical region of intermediate filaments identical to the NF-L gene [36]. The NF-M gene also contains two introns at positions equivalent to introns I and II of the NF-L gene (Myers et al., unpublished data).

Most of the introns in intermediate filament genes are located at identical positions or at equivalent positions within a few nucleotides of each other. The eight introns in the vimentin, desmin and glial fibrillary acidic protein genes are at identical sites [9,10,21] and at least five or six of these positions have been conserved in the more distantly related keratin genes [22–27]. In contrast, the NF-L gene is interrupted by only three introns that do not delineate protein domains and that occur at positions not found in other intermediate filament genes (Fig. 3). Intron sliding, a phenomenon that occurred in several genes, including fibrinogen [42], keratin [25,26], dihydrofolate reductase and serine proteinase genes [43], could not account for the present location of the three introns in the NF-L gene, since introns I, II and III are in totally different positions and too far away from the nearest neighboring intron in other intermediate filament genes (Fig. 3). Such large shifts would have provoked major disruptions in the coding sequence.

In several gene families where the genes were derived by duplication, such as the genes encoding the globins [44], the vitellogenins [45], albumin-α-fetoprotein [46], ovalbumin [47], ACTH-proen-

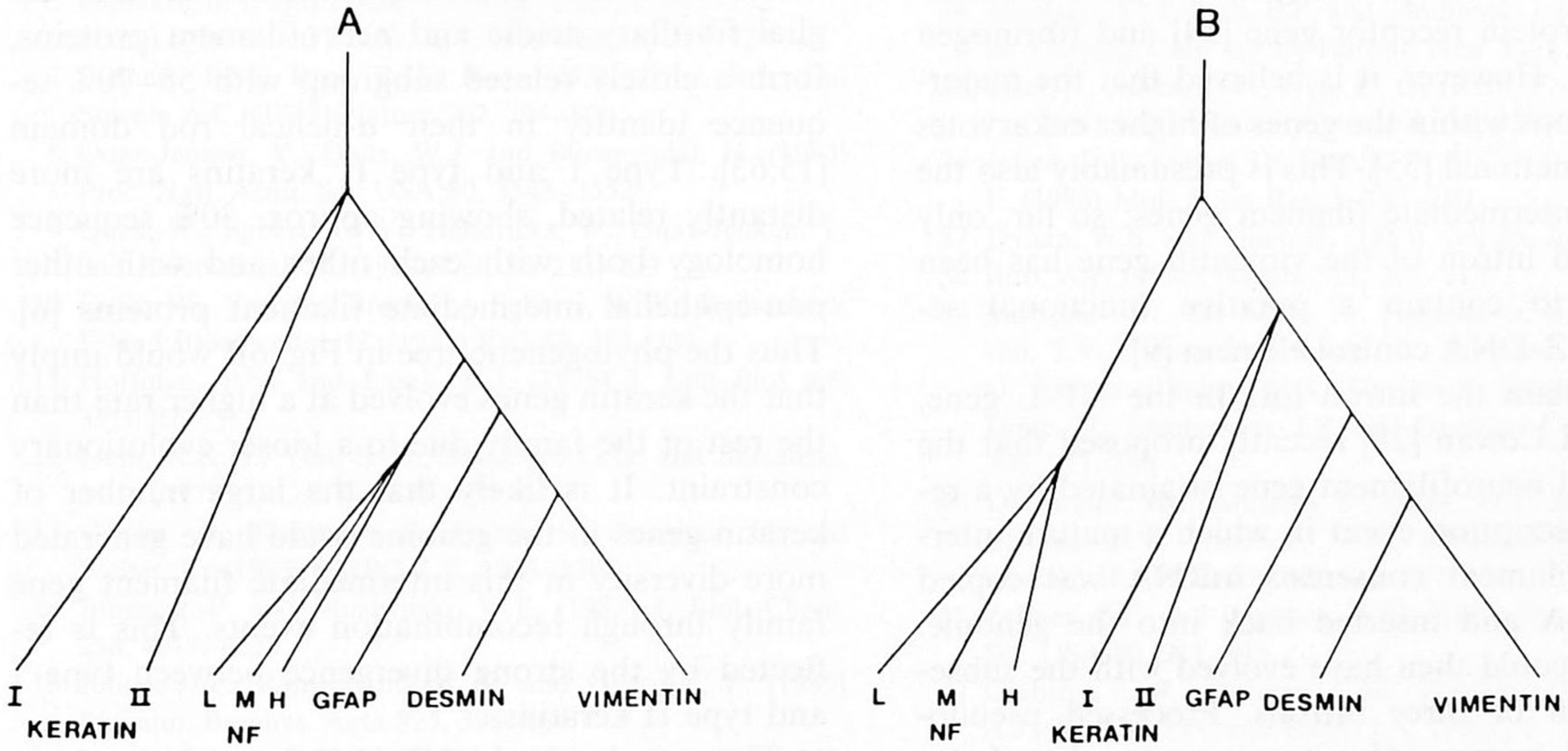

Fig. 6. Possible evolutionary models of the intermediate filament family. In panel A the branch points are calculated on the basis of the amino-acid sequence divergence in the α-helical domains of intermediate filament prpotiens, assuming a simple gene duplication event at each branch point. This model proposed by Lewis and Cowan [28] suppose an RNA-mediated transposition event at the branch point of the neurofilament gene family. In panel B the primordial neurofilament gene which could have been intronless diverged before the rest of the intermediate filament gene family. This implies that the keratin genes evolved at a higher rate than the remainder intermediate filament genes due to a looser evolutionary constraint.

kephalin [48] and the apolipoproteins [49], the intron positions are well conserved and variations in intron-exon patterns are generally explained by intron loss from an ancestral gene. The evolutionary tree drawn in Fig. 6A assumes that the keratin genes split off first. This model is based on amino-acid sequence divergence in the α-helical domains of intermediate filament proteins. Assuming a simple gene duplication event at each branch point would imply that the primordial neurofilament gene contained similar introns and that the missing introns in the neurofilament gene were lost during evolution. However, this avenue poses some problems considering the large numbers of introns that had to be lost. The elimination of a particular nonfunctional piece of DNA may be a very slow process [50]. In addition, random loss of each intron could not account for the uniqueness of the NF-L gene unless selective pressure has been responsible for the maintenance of introns in other intermediate filament genes. For example, certain introns might contain sequences that are essential for the functioning of the genes. Selective loss of introns has been proposed to explain the varying positions of introns between members of multigene families such as the rat insulin genes [51,52], actin genes [53], the low-density lipoprotein receptor gene [54] and fibrinogen genes [42]. However, it is believed that the majority of introns within the genes of higher eukaryotes are nonfunctional [55]. This is presumably also the case for intermediate filament genes; so far, only the second intron of the vimentin gene has been reported to contain a putative functional sequence, a Z-DNA control element [9].

To explain the intron loss in the NF-L gene, Lewis and Cowan [28] recently proposed that the primordial neurofilament gene originated by a reverse transcription event in which a mature intermediate filament consensus mRNA was copied into cDNA and inserted back into the genome. The gene could then have evolved with the subsequent gain of three introns. Processed pseudogenes and human *Alu* sequences are thought to have arisen by reverse transcription and direct integration into the genome [56–59]. Although this hypothesis represents a reasonable explanation, it poses some problems: first, the gene is not flanked by short direct repeats that are usually generated during integration of pseudogenes or repetitive elements; second, the gene lacks a poly(A) tract typical of mRNA-derived pseudogenes shortly downstream from the AATAAA polyadenylation signal; and third, unlike pseudogenes, the NF-L gene contains a functional promoter.

Another plausible evolutionary model is also shown in Fig. 6B. In this case the primordial neurofilament gene diverged before the keratin genes split from the rest of the intermediate filament gene family. The presence of intermediate filaments in neuronal and non-neuronal cell types of several invertebrates [60–62] suggests the emergence of a neurofilament gene in the earliest metazoa, about 700 million years ago. The primordial intermediate filament gene could have been intronless before the neurofilament branch evolved. Members of the family would have gained introns at different positions, before they each duplicated and diverged to give rise to all the different genes. Alternatively, it is possible that introns present in the primordial intermediate filament gene were lost during neurofilament gene evolution. Protein sequence divergence between the rod domains of different intermediate filaments shows that the four non-epithelial intermediate filament proteins, vimentin, desmin, the glial fibrillary acidic and neurofilament proteins, form a closely related subgroup with 55–70% sequence identity in their α-helical rod domain [15,63]. Type I and type II keratins are more distantly related, showing approx. 30% sequence homology both with each other and with other non-epithelial intermediate filament proteins [6]. Thus the phylogenetic tree in Fig. 6B would imply that the keratin genes evolved at a higher rate than the rest of the family due to a looser evolutionary constraint. It is likely that the large number of keratin genes in the genome could have generated more diversity in this intermediate filament gene family through recombination events. This is reflected by the strong divergence between type I and type II keratins.

The presence in the NF-H [36] genes of at least two intron sequences that occur at positions equivalent to the NF-L gene provides evidence that neurofilament genes evolved by duplication of a common ancestral neurofilament gene. Comparison of the neurofilament subunit compositions

of several vertebrates suggested that these duplication events occurred between 200 and 400 million years ago [64]. Determination of the complete structure of the NF-H gene is now in progress and will help to further clarify the evolution of this multigene family.

Acknowledgements

We thank Carolle St-Aubin for the preparation of the manuscript. We are grateful to R.A. Lazzarini for informative discussions and for allowing us to use the NF-M sequence for a dot-matrix analysis before publication of his manuscript. The work was supported by the Medical Research Council, U.K., and the Medical Research Council of Canada. J.-P. J. was supported by Fellowships from the Medical Research Council of Canada and the National Cancer Institute of Canada.

References

1 Lazarides, E. (1982) Annu. Rev. Biochem. 51, 219–250
2 Osborn, M. and Weber, K. (1982) Cell 31, 303–306
3 Geisler, N. and Weber, K. (1982) EMBO J. 1, 1649–1656
4 Geisler, N., Fisher, S., Van de Kerckhove, J., Plessmann, U. and Weber, K. (1984) EMBO J. 3, 2701–2706
5 Hanukoglu, I. and Fuchs, E. (1982) Cell 31, 243–252
6 Hanukoglu, I. and Fuchs, E. (1983) Cell 33, 915–924
7 Steinert, P.M., Rice, R.H., Roop, D.R., Trus, B.L. and Steven, A.C. (1983) Nature 302, 794–800
8 Quax-Jeuken, Y., Quax, W.J. and Bloemendal, H. (1983) Proc. Natl. Acad. Sci. USA 80, 3548–3552
9 Quax, W., Egberts, W.V., Hendricks, W., Quax-Jeuken, Y. and Bloemendal, H. (1983) Cell 35, 215–223
10 Quax, W., Van den Broek, L., Egberts, W.W., Ramaekers, F. and Bloemendal, H. (1985) Cell 43, 327–338
11 Hoffman, P.N. and Lasek, R.J. (1975) J. Cell Biol. 66, 351–366
12 Liem, R.K.H., Yen, S.-H., Salomon, G.D. and Shelanski, M.L. (1978) J. Cell Biol. 79, 637–645
13 Geisler, N., Kaufman, E., Fisher, S., Plessman, U. and Weber, K. (1983) EMBO J. 2, 1295–1302
14 Julien, J.-P. and Mushynski, W.E. (1983) J. Biol. Chem. 258, 4019–4025
15 Julien, J.-P., Ramachandran, K. and Grosveld, F. (1985) Biochim. Biophys. Acta 825, 398–404
16 Lewis, S.A. and Cowan, N.J. (1985) J. Cell Biol. 100, 843–850
17 Julien, J.-P. and Mushynski, W.E. (1982) J. Biol. Chem. 257, 10467–10470
18 Wong, J., Hutchison, S.B. and Liem, R.K.H. (1984) J. Biol. Chem. 259, 10867–10874
19 Geisler, N., Fisher, S., Van de Kerckhove, J., Van Damme, J.V., Plessmann, U. and Weber, K. (1985) EMBO J. 4, 57–63
20 Carden, M.J., Schlaepfer, W.W., and Lee, V.M.-Y. (1985) J. Biol. Chem. 260, 9805–9817
21 Balcarek, J.M. and Cowan, N.J. (1985) Nucleic Acids Res. 13, 5527–5543
22 Lehnert, M.F., Jorcano, J.L., Zentgraf, H., Blessing, M., Franz, J.K. and Fanke, W.W. (1984) EMBO J., 3, 3279–3287
23 Marchuk, D., McCrohon, S. and Fuchs, E. (1984) Cell 39, 491–498
24 Marchuck, D., McCrohon, S. and Fuchs, E. (1985) Proc. Natl. Acad. Sci. USA 82, 1609–1613
25 Rieger, M., Jorcano, J.L. and Franke, W.W. (1985) EMBO J. 4, 2261–2267
26 Johnson, L.D., Idler, W.N., Zhou, X-M., Roop, D.R. and Steinert, P.M. (1985) Proc. Natl. Acad. Sci. U.S.A. 82, 1896–1900
27 Steinert, P.M., Steven, A.C. and Roop, D.R. (1985) Cell 42, 411–419
28 Lewis, S.A. and Cowan, N.J. (1986) Mol. Cell Biol. 6, 1529–1534
29 Grosveld, F.G., Lund, T., Murray, E.J., Mellor, A.L., Dahl, H.H.M., and Flavell, R.A. (1982) Nucleic Acids Res. 10, 6715–6732
30 Feinberg, A.P. and Vogelstein, B. (1983) Anal. Biochem. 132, 6–13
31 Sanger, F., Coulson, A.R., Barrel, B.G., Smith, A.J.H. and Roe, B.A. (1980) J. Mol. Biol. 143, 161–178
32 Wigler, M., Sweet, R., Sim, G.K., Wold, B., Pellcier, A., Lacy, E., Maniatis, T., Silverstein, S. and Axel, R. (1979) Cell 16, 777–785
33 Southern, E. (1975) J. Mol. Biol. 98, 503–517
34 Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) Molecular Cloning, Cold Spring Harbor, New York
35 Liesi, P., Julien, J.-P., Vilja, P., Grosveld, F. and Rechardt, L. (1986) J. Histochem. Cytochem. 34, 923–926
36 Julien, J.-P., Meyer, D., Flavell, D., Hurst, J. and Grosveld, F. (1986) Mol. Brain Res. 1, 243–250
37 Dynan, W.S. and Tjian, R. (1983) Cell 35, 79–87
38 Bird, A.P. (1986) Nature 321, 209–213
39 Bartram, C.R., De Klein, A., Hagemeijer, A., Van Agthoven, T.V., Geurts Van Kessel, A., Bootsma, D., Grosveld, G., Ferguson-Smith, M.A., Davies, T., Stone, M., Heisterkamp, N., Stephenson, J.R. and Groffen, J. (1983) Nature 306, 277–280
40 Dodemont, H.J., Soriano, P., Quax, W.J., Ramaekers, F., Lenstra, J.A., Groenen, M.A., Bernardi, G. and Bloemendal, H. (1982) EMBO, J. 1, 167–171
41 Zehner, Z.E. and Paterson, B.M. (1983) Proc. Natl. Acad. Sci. USA 80, 911–915
42 Crabtree, G.R., Comeau, C.M., Fowlkes, D.M., Fornace, A.J., Malley, J.D. and Kant, J.A. (1985) J. Mol. Biol. 185, 1–19
43 Craik, C.S., Sprang, S., Fletterick, R. and Rutter, W.J. (1982) Nature 299, 180–182
44 Maniatis, T., Fritsch, E.F., Lauer, J. and Lawn, R.M. (1980) Annu. Rev. Genet. 14, 145–178
45 Wahli, W., David, I.B., Wyler, T., Weber, R. and Ryffel, G.U. (1980) Cell 20, 107–117

46 Eiferman, F.A., Young, P.R., Scott, R.W. and Tilghman, S.M. (1981) Nature 294, 713–718
47 Heilig, R., Muraskowsky, R., Kloepfer, C. and Mandel, J.L. (1982) Nucl. Acids Res. 10, 4363–4383
48 Noda, M., Teranish, Y., Takahashi, H., Toyosato, M., Notake, M., Nakanish, S. and Numa, S. (1982) Nature 297, 431–434
49 Shelley, C.S., Sharpe, C.R., Baralle, F.E. and Shoulders, C.C. (1985) J. Mol. Biol. 186, 43–51
50 Orgel, L.E. and Crick, F.H. (1980) Nature 284, 604–607
51 Lomedico, P., Rosenthal, N., Efstratiadis, A., Gilbert, W., Koldner, R., and Tizard, R. (1979) Cell 18, 545–558
52 Cordell, B., Bell, G., Tisher, E., De Noto, F.M., Ulrich, A., Pictet, R., Rutter, W.J. and Goodman, H.M. (1979) Cell 18, 533–543
53 Fyrberg, E.A., Bond, B.J., Hershey, N.D., Mixter, K.S. and Davidson, N. (1981) Cell 24, 107–116
54 Südhof, T.C., Russell, D.W., Goldstein, J.L., Brown, M.S., Sanchez-Pescador, R. and Bell, G.I. (1985) Science 228, 893–895
55 Sharp, P. (1985) Cell 42, 397–400
56 Nishioda, Y., Leder, A. and Leder, P. (1980) Proc. Natl. Acad. Sci. 77, 2806–2809
57 Hollis, G.F., Hieter, P.A., McBride, O.W., Swan, D. and Leder, P. (1982) Nature 296, 321–325
58 Van Arsdell, S.W., Denison, R.A., Bernstein, L.B. and Weiner, A.M. (1981) Cell 26, 11–17
59 Sharp, P.A. (1983) Nature 301, 471–472
60 Lasek, R.J., Krishnan, N. and Kaiserman-Abramof, J.R. (1979) J. Cell Biol. 82, 336–346
61 Bartnik, E., Osborn, M. and Weber, K. (1985) J. Cell Biol. 101, 427–440
62 Bartnik, E., Osborn, M. and Weber, K. (1986) J. Cell Biol. 102, 2033–2041
63 Geisler, N., Plessmann, U. and Weber, K. (1985) FEBS Lett. 182, 475–478
64 Lasek, J., Phillips, L., Katz, M.J. and Autilio-Gambetti, L. (1985) Ann. N.Y. Acad. Sci., Int 455, 462–478
65 Eaver, R., and Weissman, S. (1979) Nucleic Acids Res. 7, 1175–1193
66 Maxam, A., and Gilbert, W. (1980) Methods Enzymol. 65, 499–560