The EMBO Journal vol.6 no.6 pp.1617–1626, 1987

The human mid-size neurofilament subunit: a repeated protein sequence and the relationship of its gene to the intermediate filament gene family

Michael W.Myers[1,3], Robert A.Lazzarini[1], Virginia M.-Y.Lee[2], William W.Schlaepfer[2] and David L.Nelson[1,4]

[1]Laboratory of Molecular Genetics, National Institutes of Neurological Disorders and Stroke, National Institutes of Health, Bethesda MD 20892, [2]Division of Neuropathology, Department of Pathology and Laboratory Medicine, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, USA

[3]Howard Hughes Medical Institute Research Scholars Program

[4]Present address: Howard Hughes Medical Institute and Institute for Molecular Genetics, Baylor College of Medicine, Houston, TX 77030, USA

Communicated by V.Pirrotta

We report the isolation and sequence of cDNA and genomic clones for one of the two large subunits of human neurofilament, NF-M. Analysis of the sequence has allowed us to investigate the structure of the carboxy-terminal tail of this protein, and to compare it to that of the small neurofilament as well as to other intermediate filaments. The carboxy-terminal region of the protein contains a 13 amino acid proline- and serine-rich sequence repeated six times in succession. Within each repeat unit are two smaller repeats of the sequence Lys-Ser-Pro-Val. The four amino acid repeat may represent a kinase recognition site in a region of the protein that is known to be highly phosphorylated. We also note the presence of an additional heptad repeat at the extreme carboxy terminus of the protein. This region of 60 amino acids may be involved in coiled-coil interactions similar to those that facilitate the filament formation in the rod region. The human gene contains only two introns. Their positions correspond to two of the three introns found in the small neurofilament of the mouse. Thus, two of the three neurofilament genes of mammals have similar structures which are quite different from those of the other intermediate filaments. This finding suggests a common origin of the neurofilament subunits, whose evolutionary relationship to other intermediate filament genes is uncertain.

Key words: evolution/exon position/heptad repeat/phosphorylation

Introduction

The cytoskeletal architecture of cells is determined by various types of filament-forming proteins. Among these are the microtubules, actin filaments and intermediate filaments. The class of intermediate filaments contains many types, and these are expressed in both tissue-specific and developmentally regulated fashions (for review, Lazarides, 1980; Steinert *et al.*, 1985). The intermediate filament of the neuron, neurofilament, is composed of three subunit proteins. These have estimated mol. wts of 68 (NF-L), 150 (NF-M) and 200 (NF-H) kd from SDS–polyacrylamide gels. The expression of neurofilament appears to be limited to neuronal cells. Filaments formed by these subunits are found predominantly in the axons and dendrites of neurons, and are involved in their structure. The two larger subunits differ from the small neurofilament in that they contain long, highly charged and heavily phosphorylated C-terminal end extensions.

Extensive protein sequence data demonstrated homology between the three neurofilaments in the conserved 'rod' region (Geisler *et al.*, 1983,1984,1985a,b). This region is quite conserved among all the intermediate filaments, and is characterized by three regions exhibiting a heptad repeat of amino acids, with the first and fourth residue of each repeat being hydrophobic in an otherwise α-helical region. This arrangement allows the formation of the coiled-coil interactions first described by Crick (1953) (see also Cohen and Parry, 1986), and the formation of filaments. Protein sequence has not been obtained for the carboxy-terminal regions of NF-M and NF-H, although amino acid compositions have been determined (Geisler *et al.*, 1984,1985a). These carboxy-terminal regions are heavily phosphorylated at serine residues, with 9–100 phosphates present per molecule (Julien *et al.*, 1983; Carden *et al.*, 1985). The presence of phosphate contributes to overestimates of the size of NF-M and NF-H on gel electrophoresis (Kaufman *et al.*, 1984), and complicates immunocytochemical analysis of these subunits (Lee *et al.*, 1986a,b). Monoclonal antibodies directed to phosphorylated and enzymatically de-phosphorylated forms of the proteins can distinguish between the two forms *in vitro* and *in vivo* (Sternberger and Sternberger, 1982; Lee *et al.*, 1986a,b), suggesting that changes in phosphorylation state may occur in the neuron. The carboxy termini of NF-M and NF-H have been proposed as part of the structural components of the cross-bridges seen in axonal neurofilaments (Hirokawa *et al.*, 1984).

The isolation of cDNA and genomic clones for the neurofilaments is a step toward understanding their function, delineating their developmental and neuron-specific expression, exploring their relation to disease processes, and elucidating their relationship to the other intermediate filaments during evolution. Partial cDNA clones of NF-L from mouse (Lewis and Cowan, 1985), rat (Julien *et al.*, 1985) and human (D.L.Nelson *et al.*, in preparation) have been described, and more recently the entire mouse gene has been sequenced (Lewis and Cowan, 1986). The amino acid sequences predicted from the cloned DNAs have corresponded well with that provided by protein sequence data (Geisler *et al.*, 1985b). Surprisingly, the gene structure of the mouse small neurofilament showed no correspondence to the consensus structure found in the genes encoding the other intermediate filaments. Lewis and Cowan (1985) have proposed that the structure found in the NF-L gene arose after an RNA-mediated transposition event from a common ancestor of the neurofilaments, desmin, vimentin and GFAP.

We have isolated cDNA and genomic clones of the medium neurofilament, NF-M, and have determined the sequence and structure of the gene. The predicted protein for NF-M demonstrates a highly charged, highly α-helical tail region, with a proline- and serine-rich 13 amino acid sequence repeated tandemly six times. The repeated region appears to allow the otherwise helical protein to bend, and may be the major site of phosphorylation. Heptad repeats indicative of coiled-coil struc-

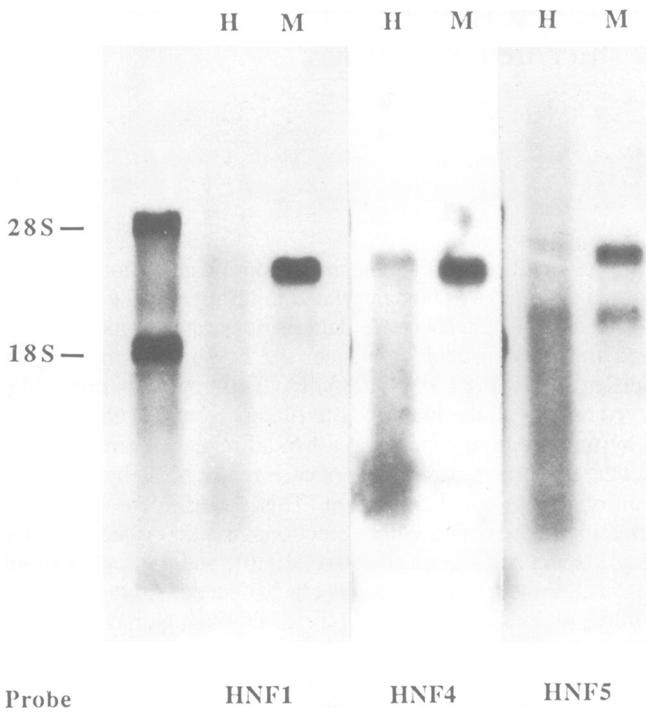

Fig. 1. Northern hybridization of cDNA clones to total RNA from human and mouse brains. Three micrograms of poly(A) selected RNA prepared from human and mouse (H and M) brains was run through a 1% agarose gel and transferred to a nylon membrane. ^{14}C-labeled ribosomal RNA was used as a size marker. Membranes were hybridized to ^{32}P-labeled DNA from the three cDNA clones indicated below. The similarity of the pattern obtained with HNF1 and HNF4 indicated that they coded differing portions of the same message.

ture have an unusual distribution in the rod segment and may also be present at the C terminus. The gene structure is more similar to that of NF-L than to the other intermediate filaments. This supports the common origin of the neurofilaments, but hypotheses other than that of an RNA-mediated transposition event leading to the evolution of neurofilaments must be considered (see Discussion).

Results

Isolation of cDNA clones for NF-M

In order to obtain cDNA clones of the human mRNAs encoding all three subunits of neurofilaments, we screened a λgt11 cDNA library prepared from basal ganglia RNA obtained from a 1-day-old infant (Kamholz *et al.*, 1986). The library was screened with antisera to all three subunits by the method of Huynh *et al.* (1984). We used a mix of polyclonal rat antisera raised to human NF-L and NF-M neurofilament subunits, and rat antisera specific for dephosphorylated versions of the two large subunits found in bovine. The screen yielded six clones from the library of ~10^6 independent cDNA inserts. These were subcloned into pUC18. On the basis of cross-hybridization, the clones could be sorted into three groups, with two clones in each. By restriction analysis, it was determined that one member of each group completely overlapped the other, so we concentrated on the longest cDNA in each group (data not shown). These clones were designated HNF1 (1200 bp), HNF4 (1000 bp), and HNF5 (1600 bp).

We took two approaches to analyze these initial cDNAs. The first was to use them as probes for Northern hybridization to brain RNA from human and mouse. As shown in Figure 1, two patterns of hybridization were obseved. HNF1 and HNF4 both detected a single band in each species RNA of a size estimated to be 3500 bp. HNF5 hybridized to two bands, one of 2200 and another of 4000. From this data, we predicted that HNF1 and HNF4 were derived from different regions of the same RNA, and that HNF5 represented a cDNA derived from a different message. The second approach was to determine the DNA sequences at the ends of each of the three clones, look at the predicted protein in the same frame as the β-galactosidase gene from the vector, and compare it to known amino acid sequence data from neurofilament proteins. We subcloned each into M13mp18, and sequenced clones in each orientation. HNF1 contained a long poly(A) tail at one end, and an unrecognizable sequence at the other. HNF4 showed virtually complete homology with the protein sequence of NF-M beginning with amino acid 122, and ending two amino acids beyond the contiguous sequence reported by Geisler *et al.* (1984). This confirmed HNF4 as a partial cDNA of the medium neurofilament subunit, and suggested that HNF1 represented a clone derived from the 3′ end of the same message. HNF5 was found to be a partial cDNA of human NF-L starting ~30 bp 5′ of the partial rat cDNA reported by Julien *et al.* (1985), and continuing through to a poly(A) tail. The detection of two RNAs in mouse and human brain with HNF5 was consistent with previous reports (Julien *et al.*, 1985; Lewis and Cowan, 1985). The sequences of HNF4 and HNF5 were completely determined and are reported elsewhere along with the chromosomal localization of the genes encoding them (D.L.Nelson *et al.*, in preparation).

After several screens of the basal ganglia and other human brain cDNA libraries (Kamholz *et al.*, 1986) using HNF1 and HNF4 as hybridization probes, we isolated >30 additional cDNA clones, two of which (HNF11 and HNF36) allowed us to overlap the region between HNF4 and HNF1. This confirmed that HNF1 was a clone from NF-M message, and allowed us to determine the sequence of all but the first 360 bp of its coding region. The positions of the four partial cDNAs relative to the mRNA are shown in Figure 2a.

Isolation and characterization of the human NF-M gene

In order to confirm the sequence of NF-M obtained from our overlapping cDNAs, to determine the sequence of the 5′ end of the message, and to investigate the structure of the gene, we isolated genomic clones from a library of human liver DNA in the λ vector EMBL3 using HNF4 as a hybridization probe. Mapping of human genomic DNA with the cDNA probes indicated that a single gene encoded NF-M (data not shown). The map of one of the clones obtained, M3, is shown in Figure 2b, and corresponds precisely with that found in human DNA. Using portions of the cDNA clones as hybridization probes on Southern transfers of M3 DNA cut with a variety of enzymes, we were able to determine that the entire gene encoding the NF-M message was located within a 6-kb region of the clone (data not shown). Fragments from this region were subcloned into M13mp18 and 19, and the sequence of each was obtained by the dideoxy method of Sanger *et al.* (1980) utilizing both the universal M13 primer and oligonucleotides synthesized to portions of the inserts.

The compiled sequence data for the NF-M gene derived from- the partial cDNAs and the genomic clones is presented in Figure 3. The sequence of the genomic DNA confirmed that derived from cDNAs, and allowed us to determine the sequence of the 5′ end of the message along with some 100 bp upstream of the ATG start site. A TATAAA sequence is found at position −61 relative to the ATG, indicating that the extent of 5′ untranslated sequence in this message is minimal. A CAAT box is not found in the upstream sequence. Two introns are located in the gene

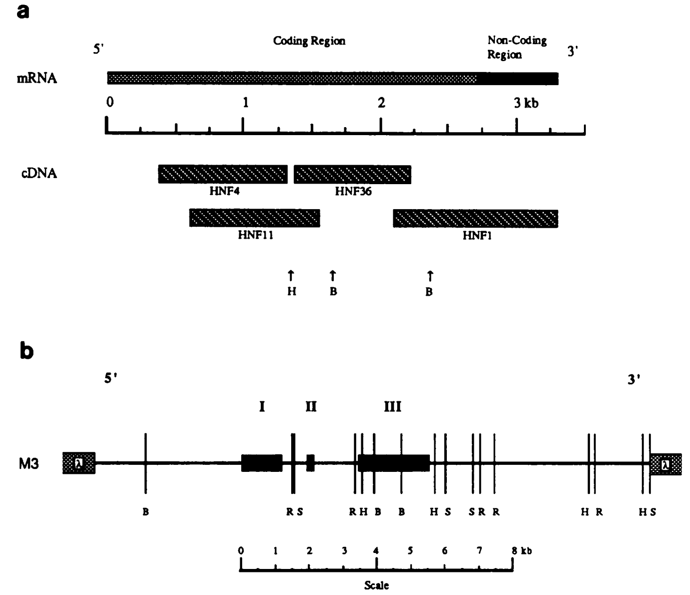

Fig. 2. **(a)** Map of cDNA clones obtained for human NF-M. Clones were mapped on the basis of common restriction sites, cross-hybridization, and sequence analysis. Sites shown are: B, *Bam*HI; H, *Hin*dIII. **(b)** Map of genomic DNA clone containing human NF-M gene. The map of one EMBL3 clone containing the human NF-M gene is displayed. The positions of the introns are idicated with black boxes, and designated I, II, and III. The direction of transcription is shown. Sites for *Bam*HI (B), *Eco*RI (R), *Hin*dIII (H), and *Sac*I (S) are shown.

as indicated in the figure. The significance of the intron locations will be discussed below. The TAA stop codon is found after nucleotide 2748, leading to the prediction of a protein of 916 amino acids. The sequence contains poly(A) addition signals at 3049 and 3235. The size of the message including the 50 base poly(A) tail found in the cDNA clone HNF1 beginning at the indicated position in Figure 3 is ~3300 bp.

The amino acid sequence of human NF-M

The amino acid sequence predicted from the nucleic acid sequence we have obtained allows the examination of the complete structure of the medium subunit of neurofilament for the first time. Comparison of the predicted amino acid sequence with that obtained from the amino-terminal half of porcine NF-M by direct protein sequencing (Geisler *et al.*, 1984) confirms that we have obtained clones for human NF-M and reveals considerable interspecies homology. The mature human protein is likely to contain 915 amino acids assuming the removal of the N-terminal methionine. The protein has a mass of ~102 kd, considerably smaller than the 150 kd estimated by SDS−PAGE but consistent with that measured by other techniques (Kaufman *et al.*, 1984).

Computer analysis of the human NF-M sequence (Figure 4) predicts α-helical conformation between amino acids 100 and 419 with a sharp decline in helical potential bracketing each side. The size of the helical region is identical to the rod regions of other intermediate filament proteins and is likewise bordered by proline residues (Weber and Geisler, 1984; Parry and Fraser, 1985; Steinert *et al.*, 1985). As in the other rod domains, the NF-M sequence can be divided into units of seven residues, or heptads, where the first ('a') and fourth ('d') residues are usually a-polar, defining the coiled-coil protein structure (Cohen and Parry, 1986). These residues are indicated in Figure 3 with circles underneath each residue in the heptad. Unlike most other intermediate filament proteins, NF-M contains no proline residues within the rod domain. In the other intermediate filament proteins, internal prolines delineate 'linker' sequences, L1, L12 and L2, which divide the rod domain into helical regions (Geisler and Weber, 1981a; Hanakoglu and Fuchs, 1982). Moreover, the heptad repeats of the NF-M rod are contiguous with a single disruption at Trp-291. This break in the coiled-coil structure is also found in other intermediate filament rod sequences (L2), and the Trp residue in this position is highly conserved. Hence, the most unusual feature of the NF-M rod compared to those of the other intermediate filaments is its apparent lack of the linker elements, L1 and L12. The only other disruption of the heptad repeats occurs near His-352, where the repeat changes phase. Similar 'skip' residues are inserted at this point in other intermediate filament rod regions, and are not thought to disrupt the integrity of the coiled-coil, although they may alter its axis and orientation (Parry and Fraser, 1985).

Comparing the NF-M sequence to the complete NF-L sequence from mouse (Lewis and Cowan, 1986) or pig (Geisler *et al.*, 1985b), we find a divergence between the two proteins beginning just after the end of the rod region continuing through the end of NF-L (Figure 5). Identical amino acids are found at only 17% of the positions aligned from amino acid 431 through the end of the small subunit. This low level of homology is in sharp contrast to the 46% amino acid homology found between the two neurofilaments in the N-terminal regions (Geisler *et al.*, 1985b). Despite the lack of strong homology between the C-terminal regions, the amino acid compositions of the two tails are very similar. Glutamate, for example, makes up 38% of NF-L and 36% of NF-M in this region.

NF-M carboxy-terminal domain: an unusual protein structure

The most striking feature of the tail region is found beginning at amino acid 614. A 13 amino acid sequence is repeated six times in a row, with slight degeneracy in the fifth and sixth repeats. These repeats are underlined and numbered in Figure 3 beginning with amino acid 614, and are also shown in Figure 6. Sequence data for this region is derived from both the cDNA HNF36 and genomic DNA clones. The same sequence was found in all four strands analyzed, thus the repeats are not the result of a cloning artifact. The sequence Lys-Ser-Pro-Val is found twice in each 13 amino acid repeat unit, and occurs once elsewhere (aa 510). Computer predictive methods suggest that the first half of each repeat unit adopts a β-turn configuration, while the carboxy-terminal half shows strong α-helical character (Figure 4). The domain containing the repeats is flanked by regions predicted as α-helical for long distances on either side (Figure 4), suggesting that the repeat domain plays a role in breaking the conformation of the protein (see Discussion).

The extreme carboxy terminus is of interest since it contains uncharged amino acids in positions consistent with a second region of potential heptad repeats starting at residue 885 and continuing through the end of the molecule for a total of nine heptads (Figure 3). However, the high threonine content, low leucine and isoleucine contents, and the appearance of polar (as well as two charged) amino acids in many of the 'a' and 'd' positions in the heptad repeat may place a strain on the capacity of this region to form coiled-coils (see Cohen and Parry, 1986).

The organization of the human NF-M gene

As shown in Figure 3, only two introns were found in the NF-M gene. They were both relatively small, with the first, 729 bp, occurring after nucleotide 1080 and the second, 1331 bp, after base 1205. The sequences of both introns have been determined and submitted to the GenBank data base, but are not shown here.

```
                 -80                 -60                 -40                 -20                                                        30
CGGGGCGGGCCCTGGCCCGGGACCAGCGCCGCGGCTATAAATGGCTGCGGCGAGGCCGGCAGAACGCTGTGACAGCCACACGCCCCAAGGCCTCCAAG ATG AGC TAC ACG TTG GAC TCG CTG GGC AAC
                                                                                                 [Met] Ser Tyr Thr Leu Asp Ser Leu Gly Asn   10

                                     60                                      90                                     120
CCG TCC GCC TAC CGG CGG GTA ACC GAG ACC CGC TCG AGC TTC AGC CGC GTC AGC GGC TCC CCG TCC AGT GGC TTC CGC TCG CAG TCG TGG TCC CGC GGC TCG CCC
Pro Ser Ala Tyr Arg Arg Val Thr Glu Thr Arg Ser Ser Phe Ser Arg Val Ser Gly Ser Pro Ser Ser Gly Phe Arg Ser Gln Ser Trp Ser Arg Gly Ser Pro   45

                                    165                                     195                                     225
AGC ACC GTG TCC TCC TCC TAT AAG CGC AGC ATG CTC GCC CCG CGC CTC GCT TAC AGC TCG GCC ATG CTC AGC TCC GCC GAG AGC AGC CTT GAC TTC AGC CAG TCC
Ser Thr Val Ser Ser Ser Tyr Lys Arg Ser Met Leu Ala Pro Arg Leu Ala Tyr Ser Ser Ala Met Leu Ser Ser Ala Glu Ser Ser Leu Asp Phe Ser Gln Ser   80

                                    270                                     300                                     330
TCG TCC CTG CTC AAC GGC GGC TCC GGA CCC GGC GGC GAC TAC AAG CTG TCC CGC TCC AAC GAG AAG GAG CAG CTG CAG GGG CTG AAC GAC CGC TTT GCC GGC TAC
Ser Ser Leu Leu Asn Gly Gly Ser Gly Pro Gly Gly Asp Tyr Lys Leu Ser Arg Ser Asn Glu Lys Glu Gln Leu Gln Gly Leu Asn Asp Arg Phe Ala Gly Tyr   115
                                                                                                 ●           ●               ●           ●

                        ▼           375                                     405                                     435
ATA GAG AAG GTG CAC TAC CTG GAG CAG CAG AAT AAG GAG ATT GAG GCG GAG ATC CAG GCG CTG CGG CAG AAG CAG GCC TCG CAC GCC CAG CTG GGC GAC GCG TAC
Ile Glu Lys Val His Tyr Leu Glu Gln Gln Asn Lys Glu Ile Glu Ala Glu Ile Gln Ala Leu Arg Gln Lys Gln Ala Ser His Ala Gln Leu Gly Asp Ala Tyr   150
             ●           ●               ●           ●               ●           ●               ●           ●               ●           ●

                                    480                                     510                                     540
GAC CAG GAG ATC CGC GAG CTG CGC GCC ACC CTG GAG ATG GTG AAC CAC GAG AAG GCT CAG GTG CAG CTG GAC TCG GAC CAC CTG GAG GAA GAC ATC CAC CGG CTC
Asp Gln Glu Ile Arg Glu Leu Arg Ala Thr Leu Glu Met Val Asn His Glu Lys Ala Gln Val Gln Leu Asp Ser Asp His Leu Glu Glu Asp Ile His Arg Leu   185
             ●           ●               ●           ●               ○           ●               ●           ●               ●           ●

                                    585                                     615                                     645
AAG GAG CGC TTT GAG GAG GAG GCG CGG TTG CGG GAC GAC ACT GAG GCG GCC ATC CGG GCG CTG CGC AAA GAC ATC GAG GAG GCG TCG CTG GTC AAG GTG GAG CTG
Lys Glu Arg Phe Glu Glu Glu Ala Arg Leu Arg Asp Asp Thr Glu Ala Ala Ile Arg Ala Leu Arg Lys Asp Ile Glu Glu Ala Ser Leu Val Lys Val Glu Leu   220
             ●           ○               ○           ●               ●           ●               ●           ●               ○           ●

                                    690                                     720                                     750
GAC AAG AAG GTG CAG TCG CTG CAG GAT GAG GTG GCC TTC CTG CGG AGC AAC CAC GAG GAG GAG GTG GCC GAC CTT CTG GCC CAG ATC CAG GCA TCG CAC ATC ACG
Asp Lys Lys Val Gln Ser Leu Gln Asp Glu Val Ala Phe Leu Arg Ser Asn His Glu Glu Glu Val Ala Asp Leu Leu Ala Gln Ile Gln Ala Ser His Ile Thr   255
             ●           ●               ●           ●               ●           ○               ●           ●               ●           ●

                                    795                                     825                                     855
GTG GAG CGC AAA GAC TAC CTG AAG ACA GAC ATC TCG ACG GCG CTG AAG GAA ATC CGC TCC CAG CTC GAA AGC CAC TCA GAC CAG AAT ATG CAC CAG GCC GAA GAG
Val Glu Arg Lys Asp Tyr Leu Lys Thr Asp Ile Ser Thr Ala Leu Lys Glu Ile Arg Ser Gln Leu Glu Ser His Ser Asp Gln Asn Met His Gln Ala Glu Glu   290
             ○           ●               ●           ●               ●           ●               ●           ●               ●

                                    900                                     930                                     960
TGG TTC AAA TGC CGC TAC GCC AAG CTC ACC GAG GCG GCC GAG CAG AAC AAG GAG GCC ATC CGC TCC GCC AAG GAA GAG ATC GCC GAG TAC CGG CGC CAG CTG CAG
Trp Phe Lys Cys Arg Tyr Ala Lys Leu Thr Glu Ala Ala Glu Gln Asn Lys Glu Ala Ile Arg Ser Ala Lys Glu Glu Ile Ala Glu Tyr Arg Arg Gln Leu Gln   325
     ●               ●           ●               ●           ●               ●           ●               ●           ●               ●

                                                                                                                                    INTRON 1
                                   1005                                    1035                                    1065                    ↓
TCC AAG AGC ATC GAG CTA GAG TCG GTG CGC GGC ACC AAG GAG TCC CTG GAG CGG CAG CTC AGC GAC ATC GAG GAG CGC CAC AAC CAC GAC CTC AGC AGC TAC CAG
Ser Lys Ser Ile Glu Leu Glu Ser Val Arg Gly Thr Lys Glu Ser Leu Glu Arg Gln Leu Ser Asp Ile Glu Glu Arg His Asn His Asp Leu Ser Ser Tyr Gln   360
     ○               ●           ●               ○           ●               ●           ●               ▼           ●               ●

                                   1110                                    1140                                    1170
GAC ACC ATC CAG CAG CTG GAA AAT GAG CTT CGG GGC ACA AAG TGG GAA ATG GCT CGT CAT TTG CGC GAA TAC CAG GAC CTC CTC AAC GTC AAG ATG GCT CTG GAT
Asp Thr Ile Gln Gln Leu Glu Asn Glu Leu Arg Gly Thr Lys Trp Glu Met Ala Arg His Leu Arg Glu Tyr Gln Asp Leu Leu Asn Val Lys Met Ala Leu Asp   395
         ●           ●               ●           ●               ●           ●               ●           ●               ○           ●

                    INTRON 2
                         ↓         1215                                    1245                                    1275
ATA GAA ATC GCT GCG TAC AGA AAA CTC CTG GAG GGT GAA GAG ACT AGA TTT AGC ACA TTT GCA GGA AGC ATC ACT GGG CCA CTG TAT ACA CAC CGA CCC CCA ATC
Ile Glu Ile Ala Ala Tyr Arg Lys Leu Leu Glu Gly Glu Glu Thr Arg Phe Ser Thr Phe Ala Gly Ser Ile Thr Gly Pro Leu Tyr Thr His Arg Pro Pro Ile   430
         ●           ●               ●           ○               ●           ●               ●

                                   1320                                    1350                                    1380
ACA ATA TCC AGT AAG ATT CAG AAA ACC AAG GTG GAA GCT CCC AAG CTT AAG GTC CAA CAC AAA TTT GTC GAG GAG ATC ATA GAG GAA ACC AAA GTG GAG GAT GAG
Thr Ile Ser Ser Lys Ile Gln Lys Thr Lys Val Glu Ala Pro Lys Leu Lys Val Gln His Lys Phe Val Glu Glu Ile Ile Glu Glu Thr Lys Val Glu Asp Glu   465

                                   1425                                    1455                                    1485
AAG TCA GAA ATG GAA GAG GCC CTG ACA GCC ATT ACA GAG GAA TTG GCC GCT TCC ATG AAG GAA GAG AAG AAA GAA GCA GCA GAA GAA AAG GAA GAG GAA CCC GAA
Lys Ser Glu Met Glu Glu Ala Leu Thr Ala Ile Thr Glu Glu Leu Ala Ala Ser Met Lys Glu Glu Lys Lys Glu Ala Ala Glu Glu Lys Glu Glu Glu Pro Glu   500

                                   1530                                    1560                                    1590
GCT GAA GAA GAA GAA GTA GCT GCC AAA AAG TCT CCA GTG AAA GCA ACT GCA CCT GAA GTT AAA GAA GAG GAA GGG GAA AAG GAG GAA GAA GAA GGC CAG GAA GAA
Ala Glu Glu Glu Glu Val Ala Ala Lys Lys Ser Pro Val Lys Ala Thr Ala Pro Glu Val Lys Glu Glu Glu Gly Glu Lys Glu Glu Glu Glu Gly Gln Glu Glu   535

                                   1635                                    1665                                    1695
GAG GAG GAA GAA GAT GAG GGA GCT AAG TCA GAC CAA GCC GAA GAG GGA GGA TCC GAG AAG GAA GGC TCT AGT GAA AAA GAG GAA GGT GAG CAG GAA GAA GGA GAA
Glu Glu Glu Glu Asp Glu Gly Ala Lys Ser Asp Gln Ala Glu Glu Gly Gly Ser Glu Lys Glu Gly Ser Ser Glu Lys Glu Glu Gly Glu Gln Glu Glu Gly Glu   570

                                   1740                                    1770                                    1800
ACA CAA GCT GAA GCT GAA GGA GAG GAA GCC GAA GCT AAA GAG GAA AAG AAA GTG GAG GAA AAG AGT GAG GAA GTG GCT ACC AAG GAG GAG CTG GTG GCA GAT GCC
Thr Glu Ala Glu Ala Glu Gly Glu Glu Ala Glu Ala Lys Glu Glu Lys Lys Val Glu Glu Lys Ser Glu Glu Val Ala Thr Lys Glu Glu Leu Val Ala Asp Ala   605

                                   1845                                    1875                                    1905
AAG GTG GAA AAG CCA GAA AAA GCC AAG TCT CCT GTG CCA AAA TCA CCA GTG GAA GAG AAA GGC AAG TCT CCT GTG CCC AAG TCA CCA GTG GAA GAG AAA GGC AAG
Lys Val Glu Lys Pro Glu Lys Ala Lys Ser Pro Val Pro Lys Ser Pro Val Glu Glu Lys Gly Lys Ser Pro Val Pro Lys Ser Pro Val Glu Glu Lys Gly Lys   640
                                --------------------------------------------------- --------------------------------------------------- ---
                                                         1                                                   2
                                   1950                                    1980                                    2010
TCT CCT GTG CCC AAG TCA CCA GTG GAA GAG AAA GGC AAG TCT CCT GTG CCG AAA TCA CCA GTG GAA GAG AAA GGC AAG TCT CCT GTG TCA AAA TCA CCA GTG GAA
Ser Pro Val Pro Lys Ser Pro Val Glu Glu Lys Gly Lys Ser Pro Val Pro Lys Ser Pro Val Glu Glu Lys Gly Lys Ser Pro Val Ser Lys Ser Pro Val Glu     675
----------------------------------------------- --------------------------------------------------- ---------------------------------------
                      3                                                  4                                            5

                                   2055                                    2085                                    2115
GAG AAA GCC AAA TCT CCT GTG CCA AAA TCA CCA GTG GAA GAG GCA AAG TCA AAA GCA GAA GTG GGG AAA GGT GAA CAG AAA GAG GAA GAA GAA AAG GAA GTC AAG
Glu Lys Ala Lys Ser Pro Val Pro Lys Ser Pro Val Glu Glu Ala Lys Ser Lys Ala Glu Val Gly Lys Gly Glu Gln Lys Glu Glu Glu Glu Lys Glu Val Lys     710
----------- ---------------------------------------------------
                                      6

                                   2160                                    2190                                    2220
GAA GCT CCC AAG GAA GAG AAG GTA GAG AAA AAG GAA GAG AAA CCA AAG GAT GTG CCA GAG AAG AAG AAA GCT GAG TCC CCT GTA AAG GAG GAA GCT GTG GCA GAG
Glu Ala Pro Lys Glu Glu Lys Val Glu Lys Lys Glu Glu Lys Pro Lys Asp Val Pro Glu Lys Lys Lys Ala Glu Ser Pro Val Lys Glu Glu Ala Val Ala Glu     745

                                   2265                                    2295                                    2325
GTG GTC ACC ATC ACC AAA TCG GTA AAG GTG CAC TTG GAG AAA GAG ACC AAA GAA GAG GGG AAG CCA CTG CAG CAG GAG AAA GAG AAG GAG AAA GCG GGA GGA GAG
Val Val Thr Ile Thr Lys Ser Val Lys Val His Leu Glu Lys Glu Thr Lys Glu Glu Gly Lys Pro Leu Gln Gln Glu Lys Glu Lys Glu Lys Ala Gly Gly Glu     780

                                   2370                                    2400                                    2430
GGA GGA AGT GAG GAG GAA GGG AGT GAT AAA GGT GCC AAG GGA TCC AGG AAG GAA GAC ATA GCT GTC AAT GGG GAG GTA GAA GGA AAA GAG GAG GTA GAG CAG GAG
Gly Gly Ser Glu Glu Glu Gly Ser Asp Lys Gly Ala Lys Gly Ser Arg Lys Glu Asp Ile Ala Val Asn Gly Glu Val Glu Gly Lys Glu Glu Val Glu Gln Glu     815

                                   2475                                    2505                                    2535
ACC AAG GAA AAA GGC AGT GGG AGG GAA GAG GAG AAA GGC GTT GTC ACC AAT GGC CTA GAC TTG AGC CCA GCA GAT GAA AAG AAG GGG GGT GAT AAA AGT GAG GAG
Thr Lys Glu Lys Gly Ser Gly Arg Glu Glu Glu Lys Gly Val Val Thr Asn Gly Leu Asp Leu Ser Pro Ala Asp Glu Lys Lys Gly Gly Asp Lys Ser Glu Glu     850

                                   2580                                    2610                                    2640
AAA GTG GTG GTG ACC AAA ACG GTA GAA AAA ATC ACC AGT GAG GGG GGA GAT GGT GCT ACC AAA TAC ATC ACT AAA TCT GTA ACC GTC ACT CAA AAG GTT GAA GAG
Lys Val Val Val Thr Lys Thr Val Glu Lys Ile Thr Ser Glu Gly Gly Asp Gly Ala Thr Lys Tyr Ile Thr Lys Ser Val Thr Val Thr Gln Lys Val Glu Glu     885
                 ●           ●               ●           ●               ●           ●               ●           ●               ●

                                   2685                                    2715                                    2745
CAT GAA GAG ACC TTT GAG GAG AAA CTA GTG TCT ACT AAA AAG GTA GAA AAA GTC ACT TCA CAC GCC ATA GTA AAG GAA GTC ACC CAG AGT GAC TAA GATTTGAGTCCA
His Glu Glu Thr Phe Glu Glu Lys Leu Val Ser Thr Lys Lys Val Glu Lys Val Thr Ser His Ala Ile Val Lys Glu Val Thr Gln Ser Asp ***                 916
 ●               ●           ○               ●           ●               ●           ●               ○           ●
                                                                                                                                        2903
TTGCAAAAGGTTAAGCCATATGACAATTTCAAAATGCATGTGATTGGCAGCTTCAAAACAGAACGGGTTCTCCCATGGGGGCTCCAGACATTGTATTTTACTTTGTGCAATATGAGGGGACTGCATGCAAGCTCAGGGTG

                                                                                                                                        3043
CTCCCTCCTCAGTCTTTGGGGGATTCAAATGCATGATATTGTATGTACCTGGGAAATTTGCCGATTTCCTAAGCTGTTGGAAGGGGGTCACTTAAGGGGGGATGTCTTGAGATGTATTATGCAAAGTACCAACTGAGCCA

                                                                                                                                        3183
AAAACAATAAATGAAACACAGAACTCAGCCTTAAGAAAGCTATATATGAATAATTATGTTTACCTCACTGGTGCATTTAAAATGGACTTTTGTTCATGGGAGAACCTCGTTGACATGCACAGTTTGCAATCTTATGTTGA
     ------

                                                                                                                    3303
TCGATGTTAAACGTCACAGCAGTACTTGCTCAATAAAGGTCATATTGGAAACATAGTCAATTGCTGAGTCTTATGTCATTTCTCTTTTTCTAATTTTTATTTATTTATTTTTATTTAGAG
                               ------
                                                       ↑
```

Fig. 3. DNA sequence of the human NF-M gene. The sequence is shown in the 5′−3′ orientation relative to the mRNA strand. Numbering of nucleotides is shown above, and of amino acids to the right-hand side, both begin with the ATG/Met as the first base or amino acid. Underlined features are the TATAA box at −60 and two poly(A) addition signals found in the 3′ non-translated region. Also underlined and numbered are the six copies of the 13 amino acid direct repeat. The positions of the two introns are shown. The arrow in the 3′ non-translated region indicates the start of 50-bp poly(A) tail found on the cDNA clone HNF1. The closed, inverted triangle above nucleotide 366 indicates the 5′ most sequence of the cDNA clone HNF4. Heptad repeats in the amino acid sequence are indicated with circles. Filled circles indicate uncharged residues in the appropriate positions, while open circles mark charged amino acids in these positions. The repeat pattern of the heptads is interrupted at Trp-291, and a 'skip' residue occurs at His-352 and is marked with an open, inverted triangle. The presence of heptad repeats in the carboxy-terminal tail is unprecedented in intermediate filaments (see text).

Both introns occur in the second α-helical domain of the rod region, positions that do not correspond to the borders of putative functional domains.

The exon structures of the human NF-M gene, those for the mouse NF-L, representatives of human type I and II keratins, and a consensus structure for the closely related desmin, vimentin and GFAP genes are shown in Figure 7. It is clear that the exon structure of NF-M is quite different from that of the other non-neuronal intermediate filaments. The positions of the first two introns correspond precisely to those reported for the mouse NF-L gene (Lewis and Cowan, 1986), leading to the conclusion that these two genes are more closely related to one another than to the other intermediate filaments. We find no intron corresponding to the third intron found in the mouse NF-L gene, however, indicating that NF-M and NF-L have not resulted from a very recent duplication event.

The point at which the two proteins diverge in the tail region can be seen on both the protein and the nucleic acid sequence levels. This is illustrated in Figure 5, comparing the sequence of human NF-M with mouse NF-L and GFAP starting with the second exon of the two neurofilaments. The NF-M and NF-L genes and proteins show 74% nucleotide and 86% amino acid identity in the regions encoded by their second exons. However, the nucleotide and amino acid homologies of NF-M and GFAP are 64 and 66% in the same regions. Comparing NF-L with GFAP gives 67 and 69% respectively. The most interesting finding is that the most conserved region among intermediate filaments, that region boxed in Figure 5, is split by an intron in the two neurofilament genes. After this second intron, the two neurofilament sequences are nearly identical for the initial 27 bp of their third exons, after which they begin to diverge. Imperfect homology continues past the position of the intron found in all other intermediate filament genes (marked in Figure 5). These observations are consistent with the presence of a common ancestor of the NF-M and NF-L genes, one with a gene structure in this region similar to the one found here. While the tails share overall nature, they do not demonstrate sequence homology, and their gene structures are not similar. The tails, then, are likely to have diverged rapidly since duplication, perhaps reflecting different functional constraints.

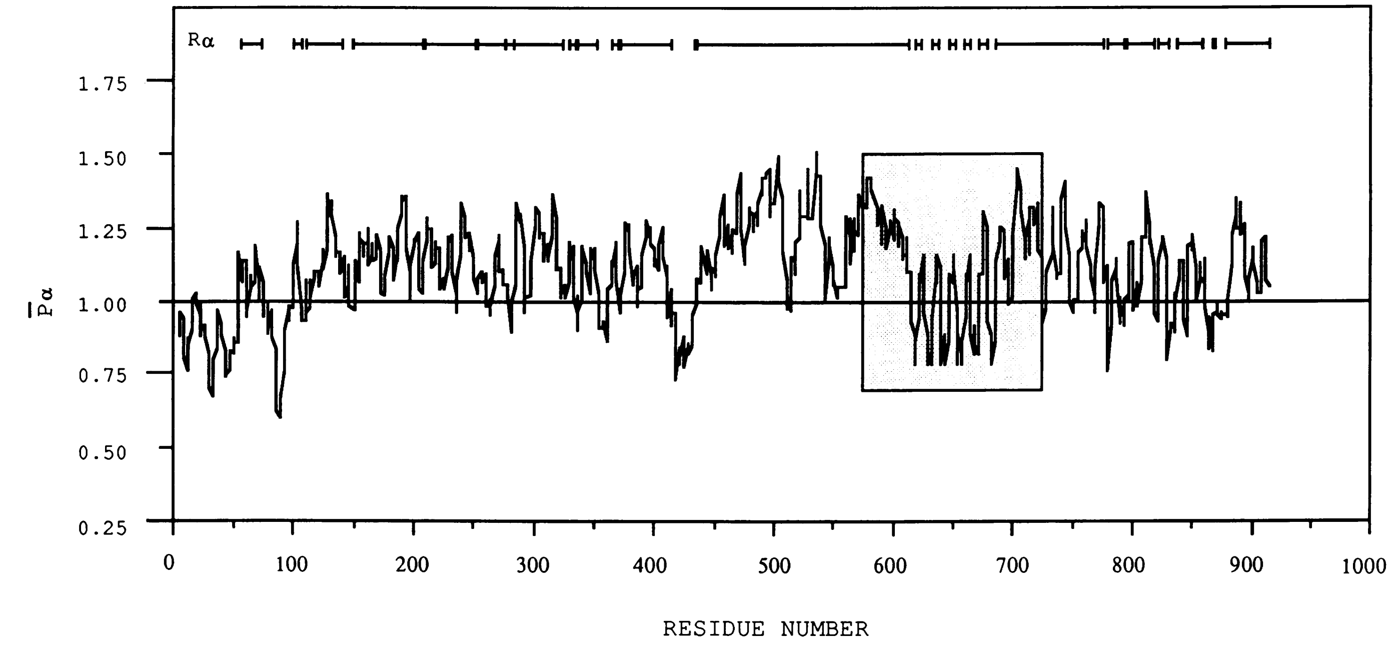

Fig. 4. Plot of probability of α-helicity for NF-M. The predicted amino acid sequence of NF-M was analyzed for α-helical regions by the computer algorithms of both Chou and Fasman (1978) and Garnier *et al.* (1978). Plotted are residue numbers of NF-M on the horizontal axis with the predicted probability of alpha helix from the Chou and Fasman method on the vertical axis. A line is drawn at the P_α of 1.00, the value at which the probability of α-helix is high. The repeat domain *c* is shaded in the plot. Areas of α-helix predicted by the Garnier method are shown as continuous lines above the Chou and Fasman plot. Cut-off value for this prediction was at the decision constant of −75, with a safety factor of 2.

```
         ▼
         D   T   I   N   K   L   E   N   E   L   R   S   T   K   S   E   M   A   R   Y   L   K   E   Y   Q   D   L   L   N   V
         D   T   I   Q   E   L   E   N   E   L   R   G   T   K   W   E   M   A   R   H   L   R   E   Y   Q   D   L   L   N   V
         E   A   L   A   R   L   E   E   E   G   Q   S   L   K   E   E   M   A   R   H   L   Q   E   Y   Q   D   L   L   N   V
NF-L    GAC ACA ATC AAC AAA CTG GAG AAT GAG CTG AGA AGC ACG AAG AGC GAG ATG GCC AGG TAC CTG AAG GAG TAC CAG GAC CTC CTC AAT GTC
NF-M    GAC ACC ATC CAG CAG CTG GAA AAT GAG CTT CGG GGC ACA AAG TGG GAA ATG GCT CGT CAT TTG CGC GAA TAC CAG GAC CTC CTC AAC GTC
GFAP    GAG GCA CTT GCT CGG CTG GAG GAG GAG GGC CAA AGC CTC AAG GAG GAG ATG GCC CGC CCA CTG CAG GAG TAC CAC GAT CTA CTC AAC GTT
        360

                                                 ▼
         K   M   A   L   D   I   E   I   A   A   Y   R   K   L   L   E   G   E   E   T   R   L   S   F   T   S   V   G   S   I
         K   M   A   L   D   I   E   I   A   A   Y   R   K   L   L   E   G   E   E   T   R   F   S   -   T   F   A   G   S   I
         K   L   A   L   D   I   E   I   A   T   Y   R   K   L   L   E   G   E   E   N   R   I   T   I   P   V   -   Q   T   F
NF-L    AAG ATG GCC TTG GAC ATC GAG ATT GCA GCT TAC AGA AAA CTC TTG GAA GGC GAA GAG ACC AGG CTC AGT TTC ACC AGC GTG GGT AGC ATA
NF-M    AAG ATG GCT CTG GAT ATA GAA ATC GCT GCG TAC AGA AAA CTC CTG GAG GGT GAA GAG ACT AGA TTT AGC --- ACA TTT GCA GGA AGC ATC
GFAP    AAG CTA GCC CTG GAC ATC GAG ATC GCC ACC TAC AGG AAA TTG CTG GAG GGC GAA GAA AAC CGC ATC ACC ATT CCT GTA --- CAG ACT TTC
                                                                                     ▲

         T   S   G   Y   S   Q   S   S   Q   V   F   G   R   S   A   Y   S   G   L   O   S   S   S   Y   L   M   S   A   R   S
         T   -   G   P   L   Y   T   H   R   P   P   -   I   T   I   S   S   K   I   Q   K   T   K   V   E   A   P   K   L   K
         S   N   L   Q   I   R   E   T   S   L   D   T   K   S   V   S   E   G   H   L   K   R   N   I   V   V   K   T   V   E
NF-L    ACC AGC GGC TAC TCT CAG AGC TCG CAG GTC TTC GGC CGT TCT GCT TAC AGT GGC TTG CAG AGC AGC TCC TAC TTG ATG TCT GCT CGC TCT
NF-M    ACT --- GGG CCA CTG TAT ACA CAC CGA CCC CCA --- ATC ACA ATA TCC AGT AAG ATT CAG AAA ACC AAG GTG GAA GCT CCC AAG CTT AAG
GFAP    TCC AAC CTC CAG ATC CGA GAA ACC AGC CTG GAC ACC AAA TCC GTG TCA GAA GGC CAC CTC AAG AGG AAC ATC GTG GTA AAG ACT GTG GAG
                             ▲                                                                                          446

                                                            ↓
NF-L    FPAYYTSHVQEEQTEVEETIEATKAEEAKDEPPSEGEAEEEEKEKEE------GEEEEGAEEE-EAAKDESEDTKEEEEGGEGEEEDTKESEEEEKKEESAGEEQVAKKD
NF-M    VQHKFVEEIIEETKVEDEKSEMEEALTAITEELAASMKEEKKEAAEEKEEEPEAEEEEVAAKKSPVKATAPEVKEEEGEKEEEEGQEEEEEEDEGAKSDQAEEGGSEKEGSS
                                                                                                                558
```

Fig. 5. Comparison of nucleic acid and amino acid sequences of NF-M, NF-L and GFAP. The region of highest homology among all intermediate filaments is compared at the nucleic acid and amino acid sequence levels for mouse NF-L (Lewis and Cowan, 1985), human NF-M (this report) and mouse glial fibrillary acidic protein (Balcarek and Cowan, 1985). Sequence begins at amino acid 361 of NF-M, which is the start of the second exon in both NF-M and NF-L, and continues through the end of NF-L (aa 559 of NF-M). Numbers correspond to amino acid residues of NF-M. GFAP sequence is compared only through the end of the rod region (aa 446 of NF-M). Boxed amino acid sequence indicates areas of high homology. Inverted closed triangles mark the positions of the two introns shared by NF-M and NF-L. Upright closed triangles mark the positions of introns in the GFAP gene. These intron positions are common to most other intermediate filament genes studied to date (see Figure 7 and text). The inverted arrow marks the position of the third intron of the mouse NF-L gene.

Discussion

Protein structure

The primary amino acid sequence of NF-M presented here allows further characterization of the carboxy-terminal domain of the protein. We find a sequence consistent with previous reports detailing the charged acidic nature of neurofilament tail regions. The most interesting feature of the protein is the repeated sequence beginning with amino acid 614. Secondary structure prediction analysis of the six tandem copies of the proline-rich 13 amino acid repeat suggest a highly flexible domain. A similar repeated region was reported by Allison *et al.* (1985) in the RNA polymerase II sequence from yeast. Here, a seven amino acid repeat was found in 26 tandem copies. The repeated sequence (Pro-Thr-Ser-Pro-Ser-Tyr-Ser), is also proline and serine rich. The authors speculate that this repeated region of RNA polymerase may be involved in interactions with other proteins, in a manner similar to that found between actin and tropomyosin in skeletal muscle (Stone and Smillie, 1978). Interactions between neurofilament and other cytoskeletal elements of the neuron (as well as between neurofilament fibers) have been observed (Leterrier *et al.*, 1982; Minami and Saicai, 1985), and the repeated region reported here may play a role in those processes. The 13 amino acid repeat unit was not found to match any known protein sequence in the NBRF protein data base, nor was any significant homology found in GenBank at the nucleic acid sequence level.

#	Lys	Ser	Pro	Val	Pro	Lys	Ser	Pro	Val	Glu	Glu	Lys	Gly
1	AAG	TCT	CCT	GTG	CCA	AAA	TCA	CCA	GTG	GAA	GAG	AAA	GCC
2	AAG	TCT	CCT	GTG	CCC	AAG	TCA	CCA	GTG	GAA	GAG	AAA	GCC
3	AAG	TCT	CCT	GTG	CCC	AAG	TCA	CCA	GTG	GAA	GAG	AAA	GGC
4	AAG	TCT	CCT	GTG	CCG	AAA	TCA	CCA	GTG	GAA	GAG	AAA	GGC
5	AAG	TCT	CCT	GTG	TCA[1]	AAA	TCA	CCA	GTG	GAA	GAG	AAA	GCC[2]
6	AAA	TCT	CCT	GTG	CCA	AAA	TCA	CCA	GTG	GAA	GAG	GCA[2]	AAG[3]

[1] Ser
[2] Ala
[3] Lys

Fig. 6. Thirteen amino acid repeat unit of NF-M. The nucleotide sequence of the six tandem copies of the 13 amino acid repeat unit beginning at base pair 1842 (Figure 3) is shown with the consensus amino acid sequence of the repeat unit. The four codons which change the amino acid sequence are designated with superscripts. The four amino acid sub-repeat, Lys-Ser-Pro-Val, is underlined for emphasis.

The position of this repeated domain in the carboxy terminus is close to that defined by peptide mapping as the site of *in vivo* phosphorylation (Julien and Mushynski, 1983). The observation that the phosphoserine residues are clustered in the protein (Zimmerman and Schlaepfer, 1986) suggests that the serine residues within the repeated domain may represent the major site of phosphorylation. The four amino acid repeat of Lys-Ser-Pro-Val is found twice per 13 residue repeat unit, and once elsewhere in the protein. Since NF-M contains 10–15 phosphoserine residues (Julien and Mushynski, 1983; Carden *et al.*, 1985), the 13 copies of Lys-Ser-Pro-Val could account for the majority of the phosphate, provided this sequence represents a recognition site for a neuronal serine kinase. We have recently shown that peptides containing the repeated sequence and phosphorylated at serine are recognized by monoclonal antibodies directed to phosphorylated epitopes of NF-M (V.M.-Y.Lee *et al.*, in

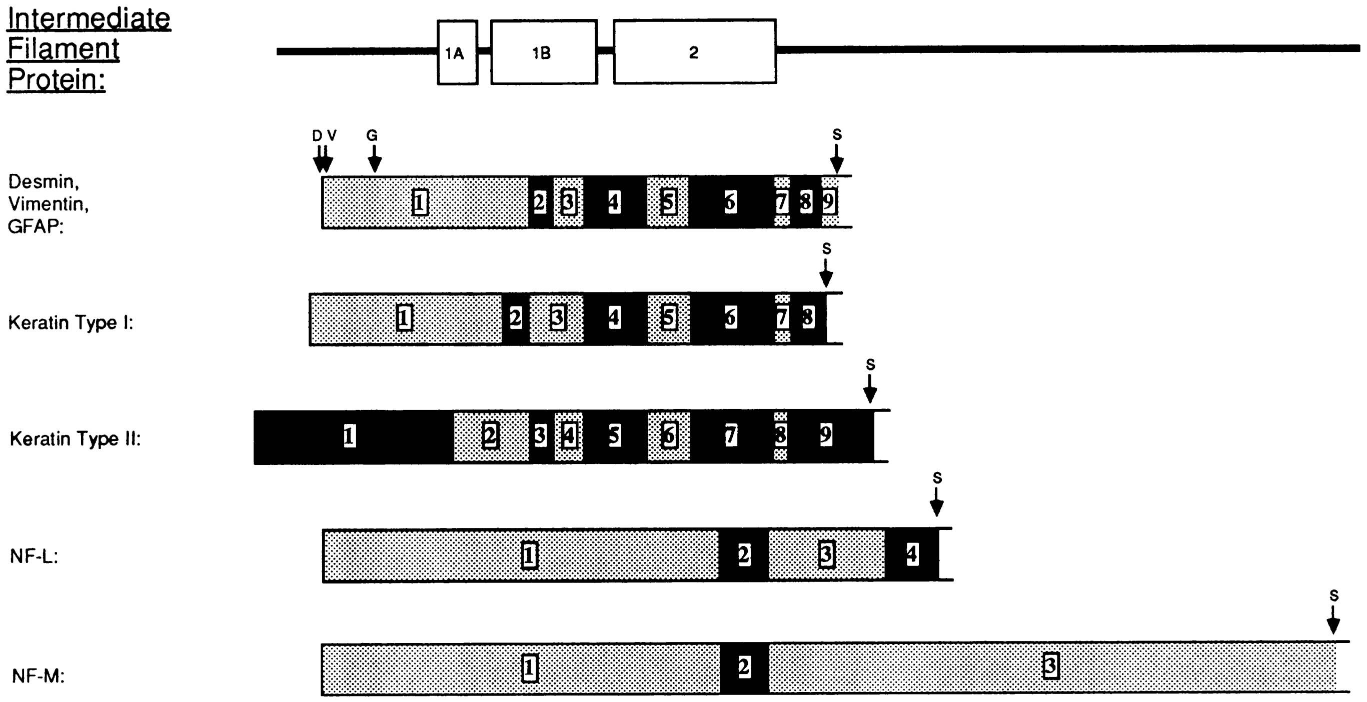

Fig. 7. Exon structures of intermediate filament genes. Gene structures for seven intermediate filament genes are shown along with a depiction of the basic structure of the protein, with the rod segments 1A, 1B and 2 designated. Breaks in stipple pattern for each gene indicate the interruption of the coding sequence by an intron, and each exon is numbered. Desmin, vimentin and GFAP have been shown by one diagram, since all three genes share intron positions. The three proteins have different lengths of amino terminal regions, however, and the start sites for each are designated by D, V and G. The keratin type I and II exon structures represent a consensus as some variability in intron position has been found in different members of these gene families. Stop codons are marked by S. Exon lengths are shown to scale. Sources for these data: hamster desmin (Quax *et al.*, 1985), hamster vimentin (Quax *et al.*, 1985), mouse GFAP (Balcarek and Cowan, 1985), human type I keratin (Marchuk *et al.*, 1984), human type II keratin (Tyner *et al.*, 1985), mouse NF-L (Lewis and Cowan, 1986), human NF-M (this report).

preparation). A kinase found in neurofilament preparations and relatively specific for neurofilament has been reported (Julien *et al.*, 1983; Pierre *et al.*, 1985). These and other kinases which phosphorylate NF-M *in vitro* have been reviewed by Nestler and Greengard (1984) and are candidates for experiments to determine whether the Lys-Ser-Pro-Val sequences can act as a substrate for phosphate addition. A search found the four amino acid repeat in 11 other proteins in the NBRF data base, primarily in four human Class II HLA sequences, three human haptoglobins and Diphtheria toxin. The surrounding sequence of each of these proteins bears little resemblance to that of NF-M. It is possible that the antigenic cross-reactivities found between phosphorylation dependent epitopes of NF-M (Lee *et al.*, 1986), NF-H (Hogue-Angeletti *et al.*, 1984) and other phosphorylated proteins such as microtubule-associated proteins 1A nd 1B (Luca *et al.*, 1986) are due to the presence of the same or similar kinase recognition sites. A partial cDNA clone of NF-M from mouse reveals conservation of the first Lys-Ser-Pro-Val repeat in human (aa 510), although eight amino acids closer to the end of the rod region (Julien *et al.*, 1986). Additionally, a partial cDNA clone of the rat large neurofilament (NF-H) shows at least four or five copies of a repeating Lys-Ser-Pro-Ala-Glu in the phosphate-containing region of the protein (Robinson *et al.*, 1986).

The consequences of phosphorylation of NF-M are profound, affecting both its physical properties and immunoreactivity (Julien and Mushynski, 1982; Sternberger and Sternberger, 1982; Bennet and DiLullo, 1985; Carden *et al.*, 1985) and appear to be limited to the carboxy-terminal tail region (Carden *et al.*, 1985). Moreover, phosphorylated forms of NF-M *in vivo* appear to be localized to axonal processes (Sternberger and Sternberger, 1982; Bennett and DiLullo, 1985), suggesting a role for phosphate addition in assembly of filaments or interaction of NF-M with other proteins composing the cytoskeletal structure of the axon. If the repeated amino acid sequence mediates interactions with other proteins and is the site of phosphate addition, the nature of any protein–protein interaction may be altered by phosphorylation state. The classic precedent for such changes is found in skeletal muscle where phosphorylation of myosin alters its interaction with actin (Huxley, 1969), and changes its conformation from a folded fibrous molecule to a straightened one (Craig *et al.*, 1983). If the repeated region is a flexible 'hinge' in the molecule, then phosphate addition to the N-terminal serines in each repeat could be a mechanism for changing the molecule's conformation. It is worth noting that phosphorylation of nuclear lamins appears to rapidly de-polymerize the filament, allowing the disassembly of the nuclear membrane during mitosis (Miake-Lye and Kirschner, 1985).

Two aspects of the heptad repeats of NF-M are interesting. The first is the presence of potential heptad repeats at the extreme carboxy-terminus of the molecule. Formation of coiled-coil structure by this region could anchor the C terminus of NF-M back into the same filament formed by the N terminus. Alternatively, the C-terminal heptads could facilitate interaction between filaments through the formation of a cross-bridge. Neurofilament cross-bridges are observed in axons, but are normally only decorated with antisera specific for the large neurofilament, NF-H (Hirokawa *et al.*, 1984), thus it seems unlikely that NF-M is involved in this process. Furthermore, as noted above, the amino acid composition of this second heptad repeat region may limit its ability to participate in coiled-coil interactions and the significance of this heptad repeat remains to be determined. The second interesting aspect of heptad repeats in NF-M, the presence of continuous repeats through the rod region, may relate both to its inability to self-assemble into filaments *in vitro* (Geisler and Weber, 1981a), and to the presence of rod region epitopes that are unique to NF-M isolated from four different species (Lee *et al.*, 1986).

Comparison of the partial porcine amino acid sequence for NF-M with that of human NF-M reported here shows a highly conserved protein, especially in the rod region, where only 1.2% of the residues have been changed. This is in contrast to the N-terminal region, where 8% of the residues show differences. Despite the porcine/human homology, comparison of the NF-M rod segment sequence with those of other intermediate filament proteins, including NF-L and NF-H, shows <60% homology (Geisler *et al.*, 1985a,b). This indicates that NF-M developed as a separate neurofilament well before the divergence of artiodactyls and primates. This is consistent with the observation of NF-M specific epitopes shared between mammals and chicken (Lee *et al.*, 1986), leading to the conclusion that NF-M diverged from the other neurofilaments at least 200 million years ago.

Evolution of the neurofilaments and other intermediate filaments

Intermediate filaments are found in all vertebrates, and immunologically cross-reactive proteins are found in most invertebrates. Indeed, a recent report details the presence of filamentous structures in higher plants which cross-react to anti-IFA, a mouse monoclonal antisera that recognizes intermediate filament epitopes non-discriminately (Dawson *et al.*, 1985). This, along with the finding that lamins A and C, structural components of the nuclear membrane, are also members of the intermediate filament family supports the thesis that intermediate filaments are an ancient class of proteins, perhaps dating back to the development of the nuclear membrane. Neurofilaments appear to have evolved as specialized intermediate filaments of nerve cells prior to the divergence of vertebrate and invertebrate animals, some 800 million years ago, since many invertebrate species display neuron-specific proteins which share epitopes and the silver staining properties of neurofilaments (Lasek *et al.*, 1985). Lasek and co-workers have studied the number and sizes of neurofilament subunits from many vertebrate and invertebrate species, and have proposed that the ancestor neurofilament gene diverged from the prototypical intermediate filament around the time of the beginning of metazoan life. This corresponds to the histological evidence for the first specialized neuronal cells. The neurofilament gene was lost from the lineage leading to arthropods, but maintained in various forms in the annelid and molluscan invertebrates as well as in the vertebrates. Subsequent gene duplications and deletions would then account for the one to four subunits found in neurofilaments of modern species.

The surprising exon structure found in the small neurofilament gene of the mouse (Lewis and Cowan, 1986) shares two introns with the human medium neurofilament gene described here. Since the neurofilaments are most similar to the desmin, vimentin and GFAP group of intermediate filaments at the amino acid sequence level through the conserved rod region, it was expected that they would share an exon structure with that group. Indeed, the keratins' exon structure, although somewhat altered and variable, is much like that of desmin, vimentin and GFAP, despite a more distant amino acid homology (Figure 7). That the exon structure of the neurofilaments should be entirely different from that of any other known intermediate filament genes led Lewis and Cowan to propose that the neurofilaments arose after the divergence of the keratins from the common ancestor of desmin, vimentin and GFAP via an RNA-mediated transposition event. This removed all the introns found in the current non-neuronal intermediate filament genes. Over time additional introns were added and the three current neurofilament subunits of mammals arose by gene duplication. Such RNA mediated transposition events are thought to be the mechanism for the production of processed pseudo-genes, which are commonly found in the genome (Karin and Richards, 1982), and have led to the presence of an active gene in at least one instance (Soares *et al.*, 1985). We have found that the NF-M gene has two introns in precisely the same positions as the first two introns of the NF-L gene, and that sequence homology between NF-M and NF-L is shared well into the third exon of each. Divergence after the beginning of the third exon has been extensive, with NF-L containing a third intron in the tail that is not shared by NF-M. If the common ancestor of these two genes was created by an RNA-mediated transposition, it must have acquired the two shared introns prior to its duplication forming NF-L and NF-M. Lewis and Cowan argue that the duplication must have occurred very soon after the transposition, since diverse phyla contain multiple neurofilament subunits. They state that the time between transposition and duplication would not have allowed similar intron placement to have developed. Our finding in the NF-M gene is at odds with this theory.

One alternative to the RNA-mediated transposition hypothesis would argue that the divergence of the neuronal intermediate filament occurred prior to that between the keratins and other intermediate filaments, with parallel or convergent selective constraints causing the neurofilaments to maintain a structure similar to that of the non-keratin intermediate filaments. In this scenario, the homology between exon structures, not the amino acid sequences, reflects the evolutionary distance. This is supported by the presence of neurofilaments and the absence of clearly defined homologues of all the vertebrate IF varieties in invertebrate species (Geisler *et al.*, 1985b; Lasek *et al.*, 1985). Within this alternative hypothesis, it could be argued that the keratins have been less constrained by selective pressure, accounting for the differences at the amino acid level. There may be strong arguments in favor of this last point. At least 20 keratin genes are found in the human genome (Marchuk *et al.*, 1984). Such a large number of copies may allow each gene to drift in sequence without adversely affecting the overall functions of the keratins. Many gene copies could also give rise to recombination events between different keratin genes similar to those described in the histocompatibility locus (Hood *et al.*, 1983), generating more diversity. The recent observations of Powell *et al.* (1986) and Ray Chaudhury *et al.* (1986) that the genes encoding keratin types I and II occur in clustered arrays supports the possibility of recombination by unequal cross-over as in other tandemly arrayed genes. Moreover, the keratins contribute to a variety of structures, hair, skin, nails, quills, horns, etc., that are directly involved in interactions of the organism with its environs. A diversity of keratin-derived structures may be a feature for which selective pressure could be quite strong. Finally, each type I keratin interacts with a specific (or a small repertoire of) type II keratins, forming heteropolymers. This is potentially a much less constraining situation than that for the desmin, vimentin and GFAP, each of which must form filaments with themselves. The requirement of homopolymerization is likely to be less tolerant of mutations in the protein, and may explain the relatively close amino acid homologies of desmin, vimentin and GFAP.

Unfortunately, finding an exon structure in the medium neurofilament that is similar to that of the small subunit does not resolve this issue. Either of the hypotheses discussed above would be consistent with similar gene structures among all three subunits of neurofilament, since they share many properties and are almost certainly derived from a common ancestor by gene duplication. However, the time since divergence of the prototype neurofilament relative to that of the keratins is the central issue, and the presence of neurofilament in the earliest metazoa provides a compelling argument that the keratins may have diverged more recently, accounting for their shared gene structure. The gene structures of the large subunit of neurofilament and of the lamins (Fisher *et al.*, 1986; McKeon *et al.*, 1986) (possibly the oldest of the intermediate filaments) may resolve this evolutionary conundrum.

Materials and methods

cDNA library screening

λgt11 plaques were screened for the presence of neurofilament epitopes essentially as described by Huynh *et al.* (1984). Forty nitrocellulose filters were made from plates containing ~ 10^5 plaques/plate, and screened with a mixture of antisera. The antisera contained polyclonal antibodies raised in rats to human NF-L, human NF-M, and dephosphorylated bovein NF-M and NF-H. Nine clones were found on the primary screen; six were eventually purified after three rounds of re-screening. Two of these were found to be the human small subunit, while the remaining four were pieces of human NF-M. These were subcloned from the phage into the plasmid vector pUC18. The library was re-screened several times for hybridizing plaques using the initial clones as probes as described by Maniatis *et al.* (1982). Nylon filters were used, and hybridized in the presence of 0.5–1 × 10^6 c.p.m./ml in 50% formamide, 5 × SSC, 0.1% SDS, 1 × Denhardt's solution and 100 mg/ml salmon sperm DNA, for 24–36 h followed by washes in 2 × SSC, 0.1% SDS (1 × 23°C and 2 × 60°C) and a final wash in 0.2 × SSC, 0.1% SDS at 60°C. Initial sequence analysis of the cDNA inserts was carried out by the method of Maxam and Gilbert (1980).

Analysis of brain RNA

RNA was prepared from brains of 18-day-old mice and from a human after autopsy by the guanidinium isothiocyanate procedure as described by Maniatis *et al.* (1982). Five micrograms of poly(A) selected RNA was run on a MOPS/formaldehyde gel and transferred to nylon. Hybridization and wash conditions were as described above.

Isolation and analysis of genomic clones

A human DNA library in the λ vector EMBL3 (Frischauf *et al.*, 1983) was screened using hybridization probes as described above for the cDNA hybridizations. DNA was prepared from the clones by the plate lysate method described in Maniatis *et al.* (1982). The insert sequences were mapped by the method of Rackwitz *et al.* (1984), using labeled oligonucleotides to identify partial digestion products containing either the left or right arm of the phage vector. Appropriate double and triple digestions were used to confirm the mapping data. Fragments from the M3 phage insert were subcloned into M13mp8, mp18 and mp19 vectors for sequence analysis. Sequence analysis was by the dideoxy chain termination method of Sanger *et al.* (1980) using both the universal primer and oligonucleotides synthesized to the derived insert sequence.

Computer analysis

Homology searches of protein (NBRF) and nucleic acid (GenBank) data bases were performed using the NAQ and PSQ programs of the NINCDS VAX (Digital Equipment Corporation) system. Protein structure analysis was performed by the Delphi program using the routines of Chou and Fasman (1978) and Garnier *et al.* (1978).

Acknowledgements

We wish to thank John Kamholz and Carmie Puckett for creating such a valuable resource in the human brain cDNA libraries, and for daily advice and direction, John Weis for the generous contribution of the human genomic library, Ward Odenwald, Bill Bellini, George Englund and Helen Engh for invaluable advice on DNA sequencing, Martin Carden for pointing out the unusual coiled-coil structure, Francesca deFerra, Janice Lees and Susan Molineaux for helpful discussions, Lynn Hudson for critical reading of the manuscript, and Mary Plumb and Blanche Lewis for excellent technical assistance. Michael Myers is a Scholar of the Howard Hughes Medical Institute Research Scholars Program.

References

Allison,L.A., Moyle,M., Shales,M. and Ingles,C.J. (1985) *Cell*, **42**, 599–610.
Balcarek,J.M. and Cowan,N.J. (1985) *Nucleic Acids Res.*, **13**, 5527–5543.
Bennett,G.S. and DiLullo,C. (1985) *J. Cell Biol.*, **100**, 1799–1804.

Carden,M.J., Schlaepfer,W.W. and Lee,V.M.-Y. (1985) *J. Biol. Chem.*, **260**, 9805–9817.

Chou,P.Y. and Fasman,G.D. (1978) *Adv. Enzymol.*, **47**, 145–148.

Cohen,C. and Parry,D.A.D. (1986) *Trends Biochem. Sci.*, **11**, 245–248.

Craig,R., Smith,R. and Kendrick-Jones,J. (1983) *Nature*, **302**, 436–439.

Crick,F.H.C. (1953) *Acta Crystallogr.*, **6**, 685–688.

Dawson,P.J., Hulme,J.S. and Lloyd,C.W. (1985) *J. Cell Biol.*, **100**, 1793–1798.

Fisher,D.Z., Chaudhary,N. and Blobel,G. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 6450–6454.

Garnier,J., Osguthorpe,D.J. and Robson,B. (1978) *J. Mol. Biol.*, **120**, 97–120.

Geisler,N. and Weber,K. (1981a) *J. Mol. Biol.*, **151**, 565–571.

Geisler,N. and Weber,K. (1981b) *Proc. Natl. Acad. Sci. USA*, **78**, 4120–4123.

Geisler,N., Kaufmann,E., Fischer,S., Plessmann,U. and Weber,K. (1983) *EMBO J.*, **2**, 1295–1302.

Geisler,N., Fischer,S., Vandekerckhove,J., Plessmann,U. and Weber,K. (1984) *EMBO J.*, **3**, 2701–2706.

Geisler,N., Fischer,S., Vandekerckhove,J., Van Damme,J., Plessmann,U. and Weber,K. (1985a) *EMBO J.*, **4**, 57–63.

Geisler,N., Plessmann,U. and Weber,K. (1985b) *FEBS Lett.*, **182**, 475–478.

Glenner,G.G. and Wong,C.W. (1984) *Biochem. Biophys. Res. Commun.*, **120**, 885–890.

Hanakoglu,I. and Fuchs,E. (1982) *Cell*, **31**, 243–252.

Hirokawa,N., Glicksman,M.A. and Willard,M.B. (1984) *J. Cell Biol.*, **98**, 1523–1536.

Hogue-Angeletti,R., Trojanowski,J.Q., Carden,M.J., Schlaepfer,W.W. and Lee,V.M.-Y. (1985) *J. Cell. Biochem.*, **27**, 181–187.

Hood,L., Steinmetz,M. and Malissen,B. (1983) *Annu. Rev. Immunol.*, **1**, 529–568.

Huxley,H.E. (1969) *Science*, **164**, 1356–1366.

Huynh,T.V., Young,R.A. and Davis,R.W. (1984) In Glover,D. (ed.), *DNA Cloning Techniques: A Practical Approach.* IRL Press, Oxford.

Johnson,L.D., Idler,W.W., Zhou,X.-M., Roop,D.R. and Steinert,P.M. (1985) *Proc. Natl. Acad. Sci.*, **82**, 1896–1900.

Julien,J.-P. and Mushynski,W.E. (1982) *J. Biol. Chem.*, **257**, 10467–10470.

Julien,J.-P. and Mushynski,W.E. (1983) *J. Biol. Chem.*, **258**, 4019–4025.

Julien,J.-P., Smaluk,G.D. and Mushynski,W.E. (1983) *Biochim. Biophys. Acta*, **775**, 25–31.

Julien,J.-P., Ramachandran,K. and Grosveld,F. (1985) *Biochim. Biophys. Acta*, **825**, 398–404.

Julien,J.-P., Meyer,D., Flavell,D., Hurst,J. and Grosveld,F. (1986) *Mol. Brain Res.*, **1**, 243–250.

Kamholz,J., deFerra,F., Puckett,C. and Lazzarini,R. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 4962–4966.

Karin,M. and Richards,R.I. (1982) *Nature*, **299**, 797–802.

Kaufmann,E., Geisler,N. and Weber,K. (1984) *FEBS Lett.*, **170**, 81–84.

Lasek,R.J., Phillips,L., Katz,M.J. and Autilio-Gambetti,L. (1985) *Ann. NY Acad. Sci.*, **455**, 462–478.

Lazarides,E. (1980) *Nature*, **283**, 249–256.

Lee,V.M.-Y., Carden,M.J. and Trojanowski,J.Q. (1986a) *J. Neurosci.*, **6**, 850–858.

Lee,V.M.-Y., Carden,M.J. and Schlaepfer,W.W. (1986b) *J.Neurosci.*, **6**, 2179–2186.

Lewis,S.A. and Cowan,N.J. (1985) *J. Cell Biol.*, **100**, 843–850.

Lewis,S.A. and Cowan,N.J. (1986) *Mol. Cell Biol.*, **6**, 1529–1534.

Leterrier,J.-F., Liem,R.K.H. and Shelanskie,M.L. (1982) *J. Cell Biol.*, **95**, 982–986.

Luca,F.C., Bloom,G.S. and Vallee,R.B. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 1006–1010.

Maniatis,T., Fritsch,E.F. and Sambrook,J. (eds) (1982) *Molecular Cloning. A Laboratory Manual.* Cold Spring Harbor Laboratory, New York.

Marchuk,D., McCrohon,S. and Fuchs,E. (1984) *Cell*, **39**, 491–498.

Masters,C.L., Simms,G., Weinman,N.A., Multhaup,G., McDonald,B.L. and Beyreuther,K. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 4245–4249.

Maxam,A.M. and Gilbert,W. (1980) In Grossman,L. and Moldave,K. (eds), *Methods in Enzymology*. Academic Press, New York, Vol. 65, pp. 499–560.

McKeon,F.D., Kirschner,M.W. and Caput,D. (1986) *Nature*, **319**, 463–468.

Miake-Lye,R. and Kirschner,M.W. (1985) *Cell*, **41**, 165–175.

Miller,C.C.J., Brion,J.-P., Calvert,R., Chin,T.K., Eagles,P.A.M., Downes,M.J., Flament-Durand,J., Haugh,M., Kahn,J., Probst,A., Ulrich,J. and Anderton,B.H. (1986) *EMBO J.*, **5**, 269–276.

Minami,Y. and Saicai,H. (1985) *FEBS Lett.*, **185**, 239–242.

Nestler,E.J. and Greengard,P. (eds) (1984) *Protein Phosphorylation in the Nervous System.* Wiley, New York.

Parry,D.A.D. and Fraser,R.D.B. (1985) *Int. J. Biol. Macromol.*, **7**, 203–213.

Pierre,M., Toru-Delbauffe,D., Francon,J., Ostu,J. and Chantoux,F. (1985) *Ann. NY Acad. Sci.*, **455**, 808–811.

Powell,B.C., Cam,G.R., Fietz,M.J. and Rogers,G.E. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 5048–5052.

Quax,W., Egberts,W.V., Hendriks,W., Quax-Jeuken,Y. and Bloemendal,H. (1983) *Cell*, **35**, 215–223.

Quax,W., van den Broek,L., Egbers,W.V., Ramaekers,F. and Bloemendal,H. (1985) *Cell*, **43**, 327–338.

Rackwitz,H.-R., Zehetner,G., Frischauf,A.-M. and Leharach,H. (1984) *Gene*, **30**, 195–200.

Ray Chaudhury,A., Marchuk,D., Lindhurst,M. and Fuchs,E. (1986) *Mol. Cell Biol.*, **6**, 539–548.

Robinson,P.A., Wion,D. and Anderton,B.H. (1986) *FEBS Lett.*, **209**, 203–205.

Sanger,F., Coulson,A.R., Barrell,B.G., Smith,A.J.H. and Roe,B.A. (1980) *J. Mol. Biol.*, **143**, 161–178.

Soares,M.B., Schon,E., Henderson,A., Karathanasis,S.K., Cate,R., Zeitlin,S.,Chirgwin,J. and Efstratiadis,A. (1985) *Mol. Cell Biol.*, **5**, 2090–2103.

Steinert,P.M., Steven,A.C. and Roop,D.R. (1985) *Cell*, **42**, 411–419.

Sternberger,L.A. and Sternberger,N.H. (1982) *Proc. Natl. Acad. Sci. USA*, **80**, 6126–6130.

Stone,D. and Smillie,L.B. (1978) *J. Biol. Chem.*, **253**, 1137–1148.

Tyner,A.L., Eichman,M.J. and Fuchs,E. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 4683–4687.

Weber,K. and Geisler,N. (1984) In Levine,A.J., Vande Woude,G.F., Topp,W.C. and Watson,J.D. (eds), *Cancer Cells. The Transformed Phenotype.* Cold Spring Harbor Laboratory, New York.

Zimmerman,U.-J.P. and Schlaepfer,W.W. (1986) *Biochemistry*, **25**, 3533–3536.

Received on December 15, 1986; revised on March 20, 1987