

Structure and evolutionary origin of the gene encoding mouse NF-M, the middle-molecular-mass neurofilament protein

Efrat LEVY¹, Ronald K. H. LIEM², Peter D'EUSTACHIO¹ and Nicholas J. COWAN¹

Departments of ¹Biochemistry and ²Pharmacology, New York University Medical Center

(Received November 26, 1986/February 13, 1987) — EJB 86 1260

We describe the complete sequence of the gene encoding mouse NF-M, the middle-molecular-mass neurofilament protein. The coding sequence is interrupted by two intervening sequences which align perfectly with the first two intervening sequences in the gene encoding NF-L (the low-molecular-mass neurofilament protein); there is no intron in the gene encoding NF-M corresponding to the third intron in NF-L. Therefore, both the number of introns and their arrangement in the genes coding NF-L and NF-M contrast sharply with the number and arrangement of introns in the genes of known sequence, encoding other members of the intermediate filament multigene family (desmin, vimentin, glial fibrillary acidic protein and the acidic and basic keratins); with the exception of a single truncated keratin gene that lacks an encoded tailpiece, these genes all contain eight introns, of which at least six are placed at homologous locations. Assuming the existence of a primordial intermediate filament gene containing most (if not all) the introns found in contemporary non-neurofilament intermediate filament genes, it seems likely that an RNA-mediated transposition event was involved in the generation of an ancestral gene encoding the NF polypeptides. A combination of insertional transposition and gene-duplication events could then explain the anomalous number and placement of introns within these genes. Consistent with this notion, we show that the genes encoding NF-M and NF-L are linked.

The eucaryotic cytoskeleton is assembled from three distinct classes of structural protein: actins, which are the subunits of microfilaments; tubulins, which polymerize to form microtubules; and the intermediate filament (IF) proteins, which assemble to form characteristic 10-nm filaments. The actins, tubulins and IF subunits are each encoded by a multigene family recognizable by homologies at the protein and nucleic acid sequence level [3, 5, 6]. In the case of the multigene family encoding IFs, the relationship between individual members is relatively distant, reflecting their early evolutionary divergence from a presumptive common ancestor and a subsequent functional specialization in terms of structure and characteristic patterns of tissue-specific expression [6–9, 16, 17]. Nonetheless, all IF proteins share a conserved central α -helical region that is responsible for their assembly into characteristic coiled coils [7–9, 40].

The neurofilament polypeptides are members of the IF multigene family that are expressed exclusively in cells of neural origin. In mammals there are three subunits, with apparent molecular masses of 68 kDa, 150 kDa and 200 kDa [9, 12, 23], hereafter referred to as NF-L, NF-M and NF-H respectively. Each is present predominantly in axons [11, 35] where they are thought to be determinants of axonal calibre [13, 14] as well as structurally involved in the maintenance of neural cell shape [14, 28]. Inspection of sequence data, obtained either directly or via cloned cDNAs, shows that, with the exception of the more divergent keratins, the members of the IF multigene family have a similar degree of homology within the coiled regions that form the backbone of the filaments [8, 19, 20, 24, 31, 32]. It therefore came as a complete

surprise to discover that, while the currently characterized genes encoding non-neurofilament IF proteins (i.e. desmin, vimentin, glial filament proteins and keratins) share identical or near identical structures with respect to the placement of their introns, the gene encoding NF-L is anomalous, containing only three introns at entirely non-homologous locations [22]. The fact that the three mammalian neurofilament polypeptides are restricted in their expression to neuronal cells, and that each possesses a highly charged carboxy-terminal domain, suggests that they share a more recent common ancestor within the IF multigene family; thus, it was to be expected that the gene encoding NF-M would have a similar structure to that encoding NF-L. Here we report that, while the two genes do indeed have a similar structure, they differ in that the third intron is absent in the gene encoding NF-M. We also show that the genes encoding the NF-L and NF-M polypeptides are linked. The implications of these observations are discussed in terms of the evolution of the IF multigene family.

EXPERIMENTAL PROCEDURE

Isolation and sequencing of the gene encoding the mouse NF-M polypeptide

A 2.7-kb cDNA clone, containing sequences encoding the entire rat NF-M polypeptide [29], was used to screen [25] a mouse genomic library cloned in bacteriophage λ EMBL4. A single positively hybridizing bacteriophage plaque was purified and amplified. The extent of sequences encoding the mouse NF-M gene was determined by restriction mapping and Southern blotting [38] using the rat NF-M-encoding cDNA labelled with ³²P by nick-translation [33] as probe. Two adjacent *Eco*RI fragments, of 3.1 kb and 3.8 kb (Fig. 2)

Correspondence to N. J. Cowan, Department of Biochemistry, New York University Medical Center, 550 First Avenue, New York City, New York, USA-10016

that hybridized with this probe were subcloned into plasmid pUC8. Exonuclease *Bal31* was used as described [21] to generate a set of overlapping, deleted fragments from each subclone, which were cloned into bacteriophage M13 and sequenced by the dideoxy-chain-termination method [34].

Chromosomal assignment of the genes encoding the mouse *NF-L* and *NF-M* polypeptides

Genomic DNA, extracted from interspecies somatic hybrid cell lines and from mouse livers, was analyzed by Southern blotting [38] as described [4].

RESULTS

A recently isolated full-length cDNA clone encoding the rat *NF-M* polypeptide [29] was used to determine the copy number of mouse genomic sequences encoding *NF-M* (Fig. 1). The data show the presence of two hybridizing bands when DNA from two strains of mouse was digested with *EcoRI*, and one hybridizing band in each *TagI* digest. The same cDNA probe was used to screen a mouse genomic library cloned in bacteriophage λ . A single positively hybridizing plaque was picked, purified and amplified for further study. The region within the recombinant fragment that contained the *NF-M* gene was determined by restriction mapping and Southern blot analysis using the cloned *NF-M* rat cDNA as probe. The data thus obtained are shown in Fig. 2. The entire hybridizing region is contained within two *EcoRI* restriction fragments of 3.1 kb and 3.8 kb consistent with a single-copy gene (see Fig. 1). These two fragments were subcloned into pUC, for sequence analysis. The region sequenced is indicated in Fig. 2; the sequence data are shown in Fig. 3 together with the encoded amino acid sequence. The mouse *NF-M* gene encodes a polypeptide of 849 amino acids that is 96.5% homologous to the sequence encoded by the corresponding rat cDNA. The gene contains two intervening sequences that interrupt the coding region at amino acids 359 and 401. Alignment of the genes encoding *NF-L* and *NF-M* for maximum amino acid homology shows that the two introns present in the *NF-M* gene correspond precisely in location to the first two introns in the *NF-L* gene. Both these introns occur in regions where there is the most extensive homology between the *NF-L* and the *NF-M* polypeptides, and indeed among all the intermediate filament subunits, including the lamins of the nuclear envelope [27]. However, there is no interruption of the coding sequence at the location corresponding to the third intron in the *NF-L* gene, or at any other position further downstream. The carboxy-terminal tail region is remarkable for its high content of purine residues (>72%) which frequently occur as long uninterrupted tracts and encode an amino acid sequence that is rich (51%) in charged residues. The evolutionary significance of the purine tracts is discussed below.

Because of the overall structural similarity between the genes encoding the *NF-L* and *NF-M* polypeptides, it seemed likely that they arose from a common ancestor by a mechanism involving gene duplication. In that event the two genes might be expected to be linked. To determine the chromosomal locations of the DNA sequences in the mouse homologous to the *NF-L* and *NF-M* cDNA probes, a panel of eight somatic hybrid cell lines, carrying various combinations of mouse chromosomes, was examined by Southern blotting. The same four hybrids had retained mouse-specific DNA

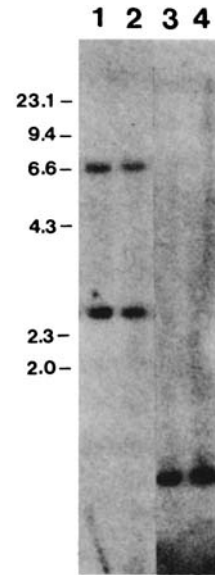


Fig. 1. A single-copy gene encodes mouse *NF-M*. Genomic DNA from two strains of mice (C57 black, tracks 1 and 3; Balb/c, tracks 2 and 4) (10 μ gm) was digested with either *EcoRI* (tracks 1 and 2) or *TagI* (tracks 3 and 4). Fragments were resolved on a 0.8% agarose gel, transferred to nitrocellulose [38] and hybridized with the cDNA probe encoding *NF-M* (see text) labelled with 32 P by nick translation [33]. The hybridization and wash conditions were as described [4]. The positions of molecular mass markers are shown at left

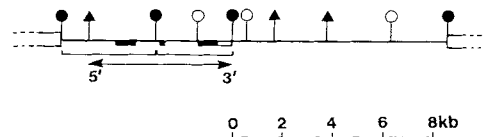


Fig. 2. Restriction map of *gNF-M*. The restriction map of the insert contained in the recombinant bacteriophage EMBL4 clone (*gNF-M*), isolated using the *NF-M*-encoding cDNA [29] as probe, is shown. (●) *EcoRI*; (▲) *BamHI*; (○) *HindIII*. Arrow delineates the region sequenced. The two *EcoRI* fragments subcloned from the λ vector are indicated by brackets. The positions of *NF-M* coding sequences and 3' non-coding sequences are indicated by filled and open boxes respectively. The direction of transcription (5'→3') is shown

fragments reactive with both probes. The only mouse chromosome present in these hybrid cell lines and absent from the others was number 14 (Table 1), indicating that the genes encoding *NF-L* and *NF-M* are both located on this chromosome in the mouse. A similar experiment was performed to determine the chromosomal assignment of the gene encoding mouse glial filament protein [2], which resulted in the localization of this IF gene to mouse chromosome 11 (data not shown).

Though the *NF-H* polypeptide appears to be under separate developmental control during brain development [36], the expression of the *NF-L* and *NF-M* polypeptides is essentially coordinate. However, comparison of upstream sequences in the two genes reveals only short stretches of regional homology. Both share a 'TATA' motif that is common to a large number of eucaryotic promoters. Two additional short upstream homologies are also evident: one (CTGCCCCCT) that appears 50 bp and 90 bp, respectively, upstream from the putative TATA box in the *NF-L* and *NF-M* genes, and a second near-perfect obverse homology

Table 1. Chromosomal mapping of NF-L and NF-M genes in somatic cell hybrids

Chromosomal contents of somatic cell hybrids were determined by karyotypic and isoenzyme analysis. Mouse chromosomes observed at sequences greater than 0.15 copy/cell are noted as present. Presence or absence of mouse DNA sequences reactive with the NF-L and NF-M cDNA probes was determined by Southern blotting, as described [4]

Hybrid	Mouse chromosomes present	NF68	NF150
BEM1-4	1,2,3, 5,6, 8, 12,13,14,15,16,17,18,19,X	+	+
MACH4B31Az3	2, 7,8, 12, 15,16,17, 19,	-	-
MACH2A2B1	1,2,3,4, 6,7,8,9,10, 12, 14,15,16,17, X	+	+
MACH2A2C3	1,2,3, 7,8,9,10, 12,13, 15,16,17, 19,X	-	-
MACH2A2H3	1, 3, 6,7,8,9,10, 12,13,14,15,16,17,18,19,X	+	+
MAE28	12, X	-	-
R44-1	17,	-	-
ECm4e	14,15,	+	+

[GGGGTGGGGTGAAGG (NF-L); CCTTCCCCCACCCC (NF-M)] between regions that lie 346 bp and 78 bp upstream from the respective TATA boxes (Fig. 3).

Amino acid comparisons of mouse NF-M (this paper), rat NF-M [29] and mouse NF-L [22] are shown in Fig. 4. The calculated molecular mass of mouse NF-M is 96 kDa. The marked difference between the authentic molecular mass and that estimated by SDS-PAGE is probably attributable to an extended helix conformation in the carboxy-terminal tailpiece, conferred by long stretches of glutamic acid residues [9, 26, 29]. Comparison of NF-M in rat and mouse reveals extensive conservation of these sequences; the coiled-coil regions (amino acids 99–244, 267–411) have changed by 1%, and the tailpiece (amino acids 412–849) by 5%, since the divergence of these two species. On the other hand, mouse NF-M and mouse NF-L are dramatically divergent: in the coiled-coil regions they are 54% homologous, similar to the homology found among all non-keratin intermediate filament coiled-coil domains [6–9, 22]. In the tail regions, however, there is no significant amino acid homology between the two sequences in spite of the great similarity in amino acid composition.

DISCUSSION

Here we report the complete structure of the gene encoding the mouse NF-M polypeptide. The data reveal the presence of only two introns, at locations homologous to the first two introns present in the gene encoding the mouse NF-L polypeptide [22]; there is no intron in the NF-M gene corresponding to the third intron present in the NF-L gene. Because the three mammalian NF proteins share similar (though not identical) patterns of tissue-specific expression and are distinct from other IF proteins by virtue of their highly acidic C-terminal tail regions, the overall structural similarity between the genes encoding the NF-L and NF-M polypeptides is not entirely surprising. Nonetheless, in accounting for the structure of contemporary IF genes we must explain both the radical difference between the structures of NF and non-NF genes, and the presence of only two introns in the gene encoding NF-M (Fig. 5). The fact that the most distantly related members of the IF multigene family (based on a comparison of homologies within the α -helical regions) share a very similar or identical gene structure has led us to argue (a) that the ancestral IF gene, from which contemporary sequences were derived, contained many (if not all) of its introns at their current locations; and (b) that the genes

encoding NF polypeptides evolved from this ancestral sequence via an mRNA-mediated transposition event [22]. Such an event would result in the insertion of an intronless IF-encoding sequence at a new genetic locus that, under appropriate circumstances (i.e. the nearby location of promoter elements and appropriate regulatory signals) could result in the expression and evolution of a novel IF sequence. Though this sequence of events may appear unlikely, a precedent for the generation of an expressed sequence following an RNA-mediated transposition event does exist in the case of rat preproinsulin gene I [37]. This gene lacks two of the introns present in the otherwise homologous preproinsulin gene II, is flanked by a direct repeat (a hallmark for chromosomal integration events) and possesses a vestigial downstream poly(A) tract that could well derive from a polyadenylated mRNA intermediate. In the case of NF-encoding sequences such characteristic hallmarks of mRNA-mediated transposition events would be lost from the contemporary genes, as a result of neutral drift during the time (about 600 million years) that has elapsed since divergence of those phyla (vertebrates, annelids and molluscs) that contain NF polypeptides [15].

Given the existence of a single-copy intronless sequence encoding a primordial NF polypeptide, three kinds of events would be required to generate the three contemporary NF polypeptide genes: the acquisition of the encoded carboxy-terminal tailpiece, the insertion of introns and at least two rounds of gene duplication. The observation that the genes encoding NF-L and NF-M contain two introns at homologous locations suggests that these introns were acquired (presumably by insertional transposition) prior to the gene duplication event. The fact that the genes encoding NF-L and NF-M are in the same linkage group (Table 1) is consistent with their origin via gene duplication. In contrast, the mouse gene encoding glial filament protein is unlinked to the two NF genes (see above), as is a cluster of at least three keratin genes that has been assigned to mouse chromosome 15 (D. Meruelo, P. D'Eustachio and M. Nesbitt, unpublished data). In addition, cosmid clones have been isolated containing up to three type I or type II keratin genes, demonstrating their very close linkage [30]. It will be of interest to see if the NF genes are also clustered in this way. Thus far it appears that distantly related members of the IF multigene family are dispersed, while gene subfamilies may be linked.

Sequences encoding the highly charged tailpieces of the NF-L and NF-M proteins contain long tracks of purines, and overall are composed of more than 72% purine residues. We previously showed that these sequences are repetitive in the mouse genome, in that a probe derived from one of them

CCGATGAAGAAAAAAGTAGAGGGGAGAACGGCCCTCGTGAGAGGCTGGGAGGGCTCTGTGTGATAAAGAACA

CGCGACCCCTACGGCCAAAAGTAAGGAAGAAAGAGACTAAGAGGTAGGGGGCCAGACGAAAGCTCAGGAAGCCAGAGAAGCTCCTCAGAGGGCCAGAGATGGCCGAGACAAGAAAGGGCC

GCTGGAAAAGGGCACTTGCCTCAAGCTGTGAACCTCAAACTGCTGGTGA AAAAGCAGAGAAGGGCTAAGCTACCGGGTGACAAGAGTCTGGAACTCAAGGGGACGCTAGGAAAAGGC

CGCCCGGGGACACCAAGTGTGGGGTCCAGTGTGGTGGCGGGCTACGTGGACGGCCGCTGAATCAGCGAAGGCTGTCTGCCCCCTTCCCCCAGCCCGGCGGTACCCCGAGTCCC

CGCTCTCGGGCCCGCTCCACGGGGCCGAGCCCTGGCCGGACAGCTGCTCCGCTATAAAGGGGCTGCGGGAGGGCCGCGAGAAGCTGTGACGCCACACCCCAAGCCCTCC

1 Met Ser Tyr Thr Leu Asp Ser Leu Gly Asn Pro Ser Ala Tyr Arg Arg Val Pro Thr Glu Thr Arg Ser Ser Phe Ser Arg Val Ser
AAG ATG AGC TAC ACG CTG GAC TCG CTG GGC AAC CCG TCC GCC TAC CGG CGC GTT CCA ACC GAG ACC ACC GAG ACC ACC GAG ACC ACC GAG ACC ACC

30 Gly Ser Pro Ser Ser Gly Phe Arg Ser Gln 40 Ser Trp Ser Arg Gly Ser Pro Ser Thr Val Ser 50 Ser Ser Tyr Thr Arg Ser Ala Val Ala
GGT TCC CCG TCC ACG GGC TTC CGC TCG CAG TCC TGG TCC CGC GGC TCG CCC AGC ACC GTG TCC TCC TCC TCC TCC TCC TCC TCC TCC TCC TCC TCC TCC

60 Pro Arg Leu Ala Tyr Ser Ser Ala Met Leu 70 Ser Ser Ala Glu Ser Ser Leu Asp Phe Ser 80 Gln Ser Ser Ser Leu Leu Asn Gly Gly Ser
CCG CGT CTC GCC TAC AGC TCG GCT ATG CTC AGC TCG GCC GAG ACC AGC CTC GAC TTC AGC CAG TCC TCG TCG TCG CTG CTC AAC GGC GGC TCC

90 Gly Gly Asp Tyr Lys Leu Ser Arg Ser Asn 100 Glu Lys Glu Gln Leu Gln Gly Leu Asn Asp Arg 110 Phe Ala Gly Tyr Ile Glu Lys Val His
GGC GGC GAC TAC AAA CTG TCC CGC TCT AAC GAG AAA GAG CAG CTG CAG GGC CTG AAC GAC CGC TTC GCC GGC TAC ATC GAG AAA GTG CAC

120 Tyr Leu Glu Gln Gln Asn Lys Glu Ile Glu 130 Glu Ala Glu Ile Gln Ala Leu Arg Gln Lys Gln Ala 140 Ser His Ala Gln Leu Gly Asp Ala Tyr
TAC TTG GAA CAA CAG AAC AAG GAG ATC GAA GCA GAG ATC CAG GCA CAG CTG CCG CAG AAG CAG GCC TCG CAC GCC CAG CTG GGT GAT GCT TAC

150 Asp Gln Glu Ile Arg Glu Leu Arg Ala Thr 160 Leu Glu Met Val Asn His Glu Lys Ala Gln 170 Val Gln Leu Asp Ser Asp His Leu Glu Glu
GAC CAG GAG ATC CGA GAG CTG CGC GCC ACC CTC GAG ATG GTG AAC CAC GAG AAG GCT CAA GTG CAG CTG GAC TCC GAT CAC TTG GAG GAA

180 Asp Ile His Arg Leu Lys Glu Arg Phe Glu 190 Glu Glu Glu Ala Arg Leu Arg Asp Asp Thr Glu Ala Ala Ile Arg Ala Leu Arg Lys Asp Ile
GAC ATC CAC CCG CTC AAG GAG CGC TTC GAG GAG GAG GCG CCG CTG CCG GAC ACC GAG GCT GCC GGC ATT CCG CCG AAA GAC ATC A

210 Glu Glu Ser Ser Met Val Lys Val Glu Leu 220 Asp Lys Lys Val Gln Ser Leu Gln Asp Glu Val 230 Phe Ala Phe Leu Arg Arg Asn His Glu Glu
GAA GAG TCG TCG ATG GTT AAG GTG GAG CTG GAC AAG AAG CTG CAG TCG CTG CAG GAT GAG GTG GCT TTC CTG CCG CGT AAT CAC GAA GAG

240 Glu Val Ala Asp Leu Leu Ala Gln Ile Gln 250 Ala Ser His Ile Thr Val Glu Arg Lys Asp Tyr 260 Tyr Leu Lys Thr Asp Ile Ser Thr Ala Leu
GAG GTG GCC GAC CTG CTG GCT CAG ATC CAG CCG TCG CAC ATC ACG GTA GAG CGC AAA GAT TAC CTG AAG ACA GAC ATC TCC ACG GCG CTG

270 Lys Glu Ile Arg Ser Gln Leu Glu Cys His 280 Ser Asp Gln Asn Met His Gln Ala Glu Glu 290 Trp Phe Lys Cys Arg Tyr Ala Lys Leu Thr
AAG GAG ATC CGC TCC CAG CTC GAG TGT CAC TCA GAC CAG AAC ATG CAC CAG GCC GAA GAG TGG TTC AAA TGC CGC TAC GCC AAG CTC ACC

300 Glu Ala Ala Glu Gln Asn Lys Glu Ala Ile 310 Arg Ser Ala Lys Glu Glu Ile Ala Glu Tyr 320 Arg Arg Gln Leu Gln Ser Lys Ser Ile Glu
GAG GCG GCC GAG CAG AAC AAG GAG GCC ATT CGC TCT GCC AAG GAA GAG ATC GCC GAG TAC CCG CGC CAG CTG CAG TCC AAG AGC ATC GAG

330 Leu Glu Ser Val Arg Gly Thr Lys Glu Ser 340 Leu Glu Arg Gln Leu Ser Asp Ile Glu Glu 350 Arg His Asn His Asp Leu Ser Ser Tyr Gln
CTC GAG TCG GTG CGA GGC ACT AAG GAG TCC CTG GAA CCG CAG CTC AGC GAC ATC GAG GAG CGC CAC AAC CAC GAC CTC AGC AGC TAC CAG

GTAGAAACCGCCGGCTGGGGCCGGCTCCGACGGCCAGGGCCGGCCCGGACACGACACTCGAGAGCGGGCCAGAGGCCCTCTTGGTCCGGCTCCCTGTGGCCCAATCCAGTCCGC

GCACGAGCTTCCGACAGCGGGTATAGCCAAAGCCCAAGTCTGTGCTCGCCCTCTCTTCGCCCAACCCACTTGCAGCACTACTTAGAAAATGCCATCTGCTCCGAAAAACTTGCTT

TCCTATGATAGATTACTGTAAAGATAAAAAGGCACATCTCGTGTGAGAAACACCCGTTTAAATAGCAGGAGCATAGATAGTCTCTGTATTCCTGCCCTCCCTCTCGTCCCC

TTCCCCCTCCCCCAACCATATCTCCACCCCTCTCTCACCCCTTACTCTCCACTCCACCCCCCGCCCCGCTCCAGCCGGATGGGGAGCTCTCAAAGCTGGCTTACCAAAA

TTTCCGGGGGGGGGGCAAAGTCCGCTAGCACCTTCTCAGATTCCTCAGTCTCCCTTATAGGGCCATAGGCTTGCAGTCTCCGTGTGTGTGTGCTAGGAAGCTGTCTGT

GATTTTGTGTCTTGTTTGAATTTTACTATCATAAAAAGAGGGGATACTAGGAGGATGTCTGCGAGGCTCATGGACTGGCGGGGAAGGGCAATTCGGGGGTGGGTAGGAAA

360 Asp Thr Ile Gln Gln Leu Glu Asn Glu Leu 370 Thr Arg Gly Thr
GGCTCGGGCAGGGAATAGCAAGTCTTGTGAAGGAGTTCTAGAGACTCCCTGTACTTTCAG GAC ACC ATC CAG CAG TCG GAA AAT GAA CTT CGG GGA ACC

380 Lys Trp Glu Met Ala Arg His Leu Arg Glu Tyr Gln Asp Leu Leu Asn Val Lys Met Ala Leu 390 Asp Ile Glu Ile Ala Ala Tyr Arg 400
AAG TCG GAA ATG GCT CGT CAT TTG CGA GAA TAC CAG GAT CTC CTT AAC GTC AAG ATG GCC CTG GAC ATC GAG ATC GCC GCG TAC AG GTA

CAGGATCTCTGACACTTGGTCCAGACCCCTACAGGCACCTGACAGGGCTGCTGCGAACACCTTCCCCACATTAAGCCAAAGCTGACCTAGTGAGCCAGGCTCAGAGCCCTGACTCCCCAG

CTCAGTTACGAACAGAACTTAAGTATTACAGATACAGGTTTACCAACTACTTACGTTTAAAGAGTGACTACGGAGAAGATCGGGAGGAGGGTGAACCTGGGAAGGGAACCAAAA

TATATTGTATGAAAAAGCCCAATATGACCAAAAGATAGTACTTTTGTAGTCTCGGGCGTCAGAAATAGTTTAAATTTTGTTC AAGCATCTGTATGCTGACAAGCAGATGAA

GTTTCATTTTAAATATTTCATGATGGTTTCCTTTTATAAAGCCAAAGGACAAAACACAGGCTGCTCTCTCTCTGTGTACAGCTTATGTTCCCATCTCTGGCTGAGACTAGATA

TAGGCACACTAGAAATAGCTGCCACTTATTTAAGTCTATGGATAATATACATGCATCAATCTGGCTAGAAATGAAATTTAATATATATTATCATATATGTAAGGCTGATACCC

TCATATCGTAGGATCAAAAAGGCTATCAGGTCATCGGTGAGAGTCTGGAGCAAACAGGAAACATATAGAAGTAGACATATAGGCAAAATAGTTTGTCTTTATCTGTAAGCTGAT

TCTATCTAGGCTCTGAGTAAGTCTCTTTCTGTGTCGACACTCAAAAGCATAACCTCTGAGGAGGAGTCACTTCTCTGTAAGCGTGTATAGCAAGAATAAAGTCACTGTCTT

Fig. 3

TTTAAAGGGAACCAACGATTTGTACCAAGAGAGCAITTCGTCACTGAACAAATTTATAATTTGGCTTACCTATATTGCTATGTAGCTATAATTTCAAATTCATCAACATTCTGACTTATG
 CTCAATAGTTTTAAITTTAAATTTGAACATTTTATTTGCTTTTAAACTCATAGAAATCTTGGCAATTCAAACGATGAGGGAGGACCCGTTTGGGGACAAATGGCTTGGCCACTAGA
 GACCATTTTACACATTCATACTACACCTACACCACTGCATGCTGCGCCAGAGCTGCATTTACTGACTGGGTGGTTAGTTGTCTACTGGGAGTGCTAGCCGCTGAGCAGAGAC
 TGCTGGCACTCTTGAGAAATCTGCCCCAGGATATGATAAACATTTGATGAAATGGCAGTAAGCACTCACTGCCCTGGTAGAAAGAGTATTTCCTTACCTGCTTCTATTATTCAA
 AGTGGCCACTACATTAAGCTTGTGGTAATGAAGATTAGAAGACCAAGAAGCTTTCTTGTATGCTCATCTCTTTGGTTCTCTCCTTAGG Lys Leu Leu Glu Gly Glu Glu
 ACC AGA TTT AGC ACA TTT TCA GCA AGC ATC ACC TGG GGG CCT CTG TAC ACA CAC CGA CAG CCC TCA GTC ACA ATA TCC AGT AAG ATT CAG AAG
 Thr Arg Phe Ser Thr Phe Ser Gly Ser Ile Thr 420 430
 410 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 Lys Ser Glu
 ACC AAA GTC GAG GCC CCC AAG CTC AAG GTC CAA CAC AAA TTT GTG GAG GAG ATC ATC GAA GAA ACT AAA GTC GAA GAT GAG AAG TCA GAA
 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 Lys Ser Glu
 440 450 460
 Met Glu Glu Thr Leu Thr Ala Ile Ala Glu Glu Leu Ala Ala Ser Ala Lys Glu Glu Lys Glu Glu Ala Glu Glu Glu Glu Glu Glu Pro
 ATG GAA GAA ACC CTC ACA GCC ATC GCA GAG GAG AAG GCA GCA TCC GCC AAA GAG GAG AAG GAA GAG GCC GAA GAA AAG GAG GAG GAA GAA
 Met Glu Glu Thr Leu Thr Ala Ile Ala Glu Glu Leu Ala Ala Ser Ala Lys Glu Glu Lys Glu Glu Ala Glu Glu Glu Glu Glu Glu Pro
 470 480 490
 Glu Ala Glu Lys Ser Pro Val Lys Ser Pro Glu Ala Lys Glu Glu Glu Glu Glu Glu Glu Glu Glu Glu Glu Glu Glu Glu Glu Glu Glu
 GAA GCC GAA AAG TCT CCC GTG AAG TCT CCT GAG GCT AAG GAA GAG GAG GAG GAA GAG GAG GAA GAG GAA GAA GAG GAA GAA GAA GAA
 Glu Ala Glu Lys Ser Pro Val Lys Ser Pro Glu Ala Lys Glu Glu Glu Glu Glu Glu Glu Glu Glu Glu Glu Glu Glu Glu Glu Glu Glu
 500 510 520
 Glu Glu Glu Glu Asp Glu Gly Val Lys Ser Asp Gln Ala Glu Glu Gly Gly Ser Glu Lys Glu Gly Ser Ser Glu Lys Asp Glu Gly Glu
 GAA GAG GAG GAA GAT CAA GGT GTC AAG TCA GAC CAG GCA GAA GAG GAG GAA GAG GAG GAA GAG GAG GAA GAG GAA GAG GAA GAG GAA
 Glu Glu Glu Glu Asp Glu Gly Val Lys Ser Asp Gln Ala Glu Glu Gly Gly Ser Glu Lys Glu Gly Ser Ser Glu Lys Asp Glu Gly Glu
 530 540 550
 Gln Glu Glu Glu Glu Gly Glu Thr Glu Ala Glu Glu Gly Glu Glu Ala Glu Ala Lys Glu Glu Lys Lys Lys Ile Glu Gly Lys Val Glu
 CAG GAA GAA GAA GAA GCA ACC GAG GCA GAA GGT GAA GGA GAG GAA GCA GAA GCT AAG GAG GAA GAA AAG ATT GAG GAA GAT GAA GGT GAG
 Gln Glu Glu Glu Glu Gly Glu Thr Glu Ala Glu Glu Gly Glu Glu Ala Glu Ala Lys Glu Glu Lys Lys Lys Ile Glu Gly Lys Val Glu
 560 570 580
 Glu Val Ala Val Lys Glu Glu Ile Lys Val Glu Lys Pro Glu Lys Ala Lys Ser Pro Met Pro Glu Lys Ser Pro Val Glu Glu Val Lys Pro
 GAA GTG GCT GTC AAG GAG GAA ATC AAG GTC GAG AAG CCT GAG GAA AAA GCC AAA TCC CCT ATG CCC AAA TCA CCC GTC GTG GAA GTA AAG CCA
 Glu Val Ala Val Lys Glu Glu Ile Lys Val Glu Lys Pro Glu Lys Ala Lys Ser Pro Met Pro Glu Lys Ser Pro Val Glu Glu Val Lys Pro
 590 600 610
 Lys Pro Glu Ala Lys Ala Gly Lys Gly Glu Gln Lys Glu Glu Glu Lys Val Glu Glu Glu Lys Lys Glu Glu Lys Lys Glu Val Thr Lys Glu Ser Pro Lys
 AAA CCA GAG GCC AAG GCC GGC AAG GGT GAG CAG AAG GAG GAA GAG AAA GTT GAG GAG AAG AAG GAA GTC ACC AAA CAA TCA CCC AAG
 Lys Pro Glu Ala Lys Ala Gly Lys Gly Glu Gln Lys Glu Glu Glu Lys Val Glu Glu Glu Lys Lys Glu Glu Lys Lys Glu Val Thr Lys Glu Ser Pro Lys
 620 630 640
 Glu Glu Lys Val Glu Lys Lys Glu Glu Lys Pro Glu Asp Val Ala Asp Lys Lys Lys Ala Glu Ser Pro Val Lys Glu Lys Ala Val Glu
 GAA GAG AAG CTG GAG AAA AAG CAG GAG AAG CCA AAA GAT GTT CCA GAT AAA AAG AAG GCC TCC CCG GTC ACC GTC AAA GAG GCT GTC GAG
 Glu Glu Lys Val Glu Lys Lys Glu Glu Lys Pro Glu Asp Val Ala Asp Lys Lys Lys Ala Glu Ser Pro Val Lys Glu Lys Ala Val Glu
 650 660 670
 Glu Val Ile Thr Ile Ser Lys Ser Val Lys Val Ser Leu Glu Lys Asp Thr Lys Glu Glu Lys Pro Gln Glu Lys Val Glu Lys Glu
 GAG CTG ATC ACC ATC ACG AAG TCG GTA AAG GTG GAG AAA GAC ACC AAA GAG GAG AAG CCG CAG CCC CAG GAG AAG GTG AAG GAG
 Glu Val Ile Thr Ile Ser Lys Ser Val Lys Val Ser Leu Glu Lys Asp Thr Lys Glu Glu Lys Pro Gln Glu Lys Val Glu Lys Glu
 680 690 700
 Lys Ala Glu Glu Glu Gly Gly Ser Glu Glu Glu Gly Ser Asp Arg Ser Pro Gln Glu Ser Lys Lys Glu Asp Ile Ala Ile Asn Gly Glu
 AAG CCA GAG GAG GAG GCG GGC ACT CAG GAG GAA T G GGT AGT GAC CGT ACG CCG CAG GAG TCC AAG AAG CAA GAC ATA GCT ATC AAT GGG GAG
 Lys Ala Glu Glu Glu Gly Gly Ser Glu Glu Glu Gly Ser Asp Arg Ser Pro Gln Glu Ser Lys Lys Glu Asp Ile Ala Ile Asn Gly Glu
 710 720 730
 Val Glu Gly Lys Glu Glu Glu Glu Gln Glu Thr Gln Glu Lys Gly Ser Gly Arg Glu Glu Glu Lys Gly Val Val Thr Asn Gly Leu Asp
 CTG GAA GGA AAA GAG CAG GAG CAG CAG CAA ACT CAG GAG AAG GCG AGT GGG AA Arg GCG GAG GAG GAG AAA GGG GTC GTC ACT AAT GGC TTA GAT
 Val Glu Gly Lys Glu Glu Glu Glu Gln Glu Thr Gln Glu Lys Gly Ser Gly Arg Glu Glu Glu Lys Gly Val Val Thr Asn Gly Leu Asp
 740 750 760
 Val Ser Pro Ala Glu Glu Lys Lys Gly Glu Asp Ser Ser Asp Asp Lys Val Val Val Thr Lys Lys Val Glu Lys Val Glu Lys Val Glu
 GTG AGC CCT GCA GAG CAG AAG AAA GGA GAG GAT A AGC AGT GAT GAT AAA GTG GTG GTC ACC AAG AAG GTA GAA AAG ATC ACC AGC GAG GGA
 Val Ser Pro Ala Glu Glu Lys Lys Gly Glu Asp Ser Ser Asp Asp Lys Val Val Val Thr Lys Lys Val Glu Lys Val Glu Lys Val Glu
 770 780 790
 Gly Asp Gly Ala Thr Lys Tyr Ile Thr Lys Ser Val Thr Val Thr Gln Lys Val Glu Glu His Glu Glu Thr Phe Glu Glu Lys Leu Val
 GGC CAT GGT GCT ACC AAA TAC ATC ACC AAA TCT GTA ACC GTC ACT ACT CAA AAG GTT GAA GAG CAT GAG CAG ACC TTT GAG GAG AAG CTG GTC
 Gly Asp Gly Ala Thr Lys Tyr Ile Thr Lys Ser Val Thr Val Thr Gln Lys Val Glu Glu His Glu Glu Thr Phe Glu Glu Lys Leu Val
 800 810 820
 Ser Thr Lys Lys Val Glu Lys Val Thr Ser His Ala Ile Val Lys Glu Val Thr Gln Gly Asp
 TCA ACT AAA AAG GTA CAA AAG GTC ACT TCA CAC GCC ATA GTC AAG GAA GTC ACC CAG GGT GAC TAAGATCCGAGTCCCGTTCGCAAAAGGTTAAGCCATA
 Ser Thr Lys Lys Val Glu Lys Val Thr Ser His Ala Ile Val Lys Glu Val Thr Gln Gly Asp
 830 840
 CGACAATTTCAAATGCGATTGACAGCTTCAAACAGAAATGGGTTCCCTCCAGGGGCTCCAGACATGTATTTCCTTTTGTGCAATATGAGGGAACGCCAAGCTCAGGGT
 A A
 GCCCCCTCTCAGTCCITGGGGGAATTC

Fig. 3. Complete sequence of the gene encoding mouse NF-M. The sequence of the mouse NF-M gene is shown. An upstream 'TATA' motif is underlined. Differences between the mouse genomic sequence and the sequence of the rat NF-M cDNA, used to screen the library, are noted; blank spaces denote sequence identity. Sequences homologous to upstream regions in the gene encoding mouse NF-L [22] are indicated by dashed lines, as is an *EcoRI* restriction site used for subcloning (see Fig. 2)

cross-hybridizes at relatively high stringency (68°C, 2 × standard saline/citrate) with a continuum of restriction fragments present in mouse genomic DNA [17], perhaps in part because this purine-rich probe bears significant homology to mouse satellite DNA [17]. Thus, if the hypothesized reverse transcript led to the primordial neurofilament gene integrated at a chromosomal locus adjacent to one of these

abundant purine-rich repetitive sequences, a portion of these sequences might then have been recruited to form the carboxy terminus of the new protein.

In any case it is clear from the unique intron/exon structure shared by the genes encoding NF-M and NF-L (Fig. 5) that these genes must derive from a common ancestral gene and that they form a distinct branch of the IF-gene family

NF-M mouse	MSYTLDSLGNPSAYRRVPTETRRSSFSRVSGSPSSGFRSQSWSRSGSPSTVS
NF-M rat	
NF-L mouse	SFAS PIFST YK YVETP VHI S RSGY TARSAYSYSYA* V**
NF-M mouse	SSYTRSAVAPRLAYSSAMLSAESSLDFSQSSSLNNGSGGDKYKLSRSNE
NF-M rat	K L
NF-L mouse	LS*****V RS SS***GSLKPSLENLDVSVQVAAIN L SI IQ
NF-M mouse	KEQLQQLNDRFAGYIEKVVHYLEQKNKEIEAEIQALRQKQASHAQLGDAYD
NF-M rat	H
NF-L mouse	A D SF R E VL GLLV HSGPSRFRAL E
NF-M mouse	QEIRELRATLEMVNHEKAQVQLDSDHLEEDIHRLKERFEEEARLRDDEA
NF-M rat	
NF-L mouse	D LAA DATN QALEGEREG TLRN QA Y VLS E A G
NF-M mouse	AIRALRKDIEESSMVKVELDKKQVSLQDEVAFLRRNHHEEVADLLAQIQ
NF-M rat	V S
NF-L mouse	RLMEA GAD AALARA E RID M I KKV I E Q I
NF-M mouse	SHITVERKDYLDKTDISTALKEIRSQLECHSDQNMHQAEWFKCRYAKLITE
NF-M rat	
NF-L mouse	AQ S MDVSS P L A D A Y KLAAK QN S FTV
NF-M mouse	AAEQNKKAIRSAKEEIAEYRRQLQSKSIELESVRGTKESLERQLSDIEER
NF-M rat	
NF-L mouse	S AK TD V A D VS S L KA TL I AC MN A K QEL DK
NF-M mouse	HNHDLSSYQDTIQLELNLGRGTMARHLREYQDMLNVKMLDIEIAAY
NF-M rat	
NF-L mouse	Q A I AN NK S S Y K
NF-M mouse	RKLLEGEETRFSTFGSITGPLYTHRQPSVTIISKIQKTKVEAPKLVQ
NF-M rat	
NF-L mouse	L FT SGYSQSS VFGRSAYSGLQSSSYLMSARSFP
NF-M mouse	KFVEEIIIEETKVEDEKSEMEETTATAEELAAASAKEEKEEAEKEEPEEA
NF-M rat	DA V
NF-L mouse	AYTSHVQ EQT V ETIEATKAEAAKD PPSEGEA E K KE G E E
NF-M mouse	EKSPVKSPEAKEEEEEEGEKEEEEEEGQEEEEEEDEGVKSDQAEEGGSEK
NF-M rat	*
NF-L mouse	GAEEEAAKDES DTKE E GG GEE DTK SE EE KEESAGEEQVAKK
NF-M mouse	SSEKDEGEQEEEEGETEAEGEGEEAEAKKEKKIEGKVEEVAVKKEIKVEK
NF-M rat	* T M I
NF-L mouse	KDJ
NF-M mouse	PEKAKSPMPKSPVEEVKPKPEAKAGCKGEQKEEKEVEEKKVETKESPKEE
NF-M rat	V D *
NF-L mouse	KVEKKEEKPKDVADKKAESPVKEKAVEEVIITISKSVKVSLEKDTKEEKP
NF-M rat	P M T
NF-M mouse	QPQEKVKEKAEEEGGSEEGSDRSPQESKKEDIAINGEVEGKEEEEQETQ
NF-M rat	Q VG K
NF-M mouse	EKSGREEKGVVNTGLDVSPEAEKKGEDSDDKVVVTKKVEKITSEGGD
NF-M rat	Q R
NF-M mouse	GATKYITKSVTVIQKVEEHEETFEKLVSTKKEVKTSHAIKVEVTQCD
NF-M rat	

Fig. 4. Comparison of NF-M and NF-L amino acid sequences. The amino acid sequences of mouse NF-M (this paper), rat NF-M [29] and mouse NF-L [22] are compared. Blank spaces denote homology; asterisks have been introduced to denote 'deletions' so as to maximize overall homology. The location of the carboxy terminus of NF-L is indicated by a square bracket

evolutionary tree. Following the triplication of the putative primordial NF-encoding gene, the regions encoding the carboxy-terminal tailpieces of the NF-M and NF-L proteins (Fig. 4) must have diverged to such an extent that there is no remaining homology between them over and above that which can be accounted for solely on the basis of the purine-rich nature of both sequences (Figs 3, 4). Furthermore, the gene encoding NF-L has a third intron in this region [22] not present in the gene encoding NF-M, which must either have been lost from the NF-M gene or inserted into the NF-L gene subsequent to the triplication of the primordial gene. Because of the tendency of repeat-containing sequences to drift, it is surprising to note the stability of the sequences encoding the carboxy-terminal regions of NF-M: the amino acid sequences of mouse and rat (Fig. 4) are 95% identical over this region. Thus, the tailpiece is almost as conserved as the coiled-coils, and must therefore be under a similar (though somewhat lesser) degree of selective pressure.

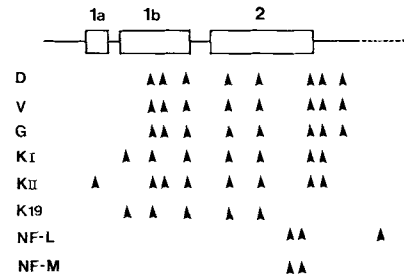


Fig. 5. Comparison of intron locations in intermediate filament genes of known structure. The α -helical domains are boxed (coils 1a, 1b, and 2); the extended carboxy-terminal tail region characteristic of the NF polypeptides is shown (---). D = hamster desmin [31], V = hamster vimentin [32], G = mouse glial fibrillary acidic protein [2], KI = human type I cytokeratin [24], KII = human type II cytokeratin [39], K19 = a bovine type I cytokeratin [1], mouse NF-L [22] and mouse NF-M (this paper). The position of the final intron varies somewhat among the keratin gene family members (see for example [1, 18]) but this can be explained by the fact that this intron interrupts the sequences encoding the tailpiece, a region of great variability. Similarly, the absence of the final two introns in K19 is not surprising, as this keratin is truncated and terminates at the end of coil 2 [1]

REFERENCES

- Bader, B. L., Magin, T. M., Hatzfeld, M. & Franke, W. W. (1986) *EMBO J.* 5, 1865–1875.
- Balcarek, J. M. & Cowan, N. J. (1985) *Nucleic Acids Res.* 13, 5527–5543.
- Cleveland, D. W. & Sullivan, K. F. (1986) *Annu. Rev. Biochem.* 54, 331–365.
- D'Eustachio, P., Kristensen, T., Wetsel, R. A., Riblet, R., Taylor, B. A. & Tack, B. F. (1987) *J. Immunol.*, in the press.
- Engel, J., Gunning, P. & Kedes, L. (1982) in *Muscle development, molecular and cellular control* (Pearson, M. L., & Epstein, H., eds) pp. 107–118, Cold Spring Harbor Laboratory, NY.
- Fuchs, E. & Hanukoglu, I. (1983) *Cell* 34, 332–334.
- Geisler, N. & Weber, K. (1982) *EMBO J.* 1, 1649–1656.
- Geisler, N., Kaufman, E., Fisher, S., Plessmann, U. & Weber, K. (1983) *EMBO J.* 2, 1295–1302.
- Geisler, N., Fisher, S., Vandekerckhove, J., Plessmann, U. & Weber, K. (1984) *EMBO J.* 3, 2701–2706.
- Hanukoglu, I. & Fuchs, E. (1982) *Cell* 31, 243–252.
- Hirokawa, N., Glicksman, M. A. & Willard, M. (1984) *J. Cell Biol.* 98, 1523–1536.
- Hoffman, P. N. & Lasek, R. J. (1975) *J. Cell Biol.* 66, 351–366.
- Hoffman, P. N., Griffin, J. W. & Price, D. L. (1984) *J. Cell Biol.* 99, 705–714.
- Lasek, R. J., Oblinger, M. H. & Drake, P. F. (1983) *Cold Spring Harbor Symp. Quant. Biol.* 48, 731–744.
- Lasek, R. J., Phillips, L., Katz, M. J. & Autilio-Gambetti, L. (1985) *Ann. N.Y. Acad. Sci.* 455, 462–478.
- Lazarides, E. (1980) *Nature (Lond.)* 283, 249–256.
- Lazarides, E. (1980) *Annu. Rev. Biochem.* 51, 219–250.
- Lehnert, M. E., Jorcano, J. L., Zentgraf, H., Blessing, M., Franz, J. K. & Franke, W. W. (1984) *EMBO J.* 3, 3279–3287.
- Lewis, S. A., Balcarek, J. M., Krek, V., Shelanski, M. & Cowan, N. J. (1984) *Proc. Natl. Acad. Sci. USA* 81, 2743–2746.
- Lewis, S. A. & Cowan, N. J. (1985) *J. Cell Biol.* 100, 843–850.
- Lewis, S. A., Gilmartin, M. E., Hall, J. L. & Cowan, N. J. (1985) *J. Mol. Biol.* 182, 11–20.
- Lewis, S. A. & Cowan, N. J. (1986) *Mol. Cell Biol.* 6, 1529–1534.
- Liem, R. K. H., Yen, S.-H., Salomon, G. D. & Shelanski, M. L. (1978) *J. Cell Biol.* 79, 637–645.
- Marchuk, D., McCrohan, S. & Fuchs, E. (1984) *Cell* 39, 491–498.

25. Maniatis, T., Hardison, E., Lacy, E., Laver, J., O'Connell, D., Quon, D., Sun, G. K. & Efstratiadis, A. (1978) *Cell* 15, 686–701.
26. Maxfield, F. R., Alter, J. E., Taylor, G. T. & Scheraga, H. A. (1975) *Macromolecules* 8, 479–491.
27. McKeon, F. D., Kirshner, M. W. & Caput, P. (1986) *Nature (Lond.)* 319, 463–468.
28. Morris, J. R. & Lasek, R. J. (1982) *J. Cell Biol.* 92, 192–198.
29. Napolitano, E. W., Chin, S. M., Colman, D. R. & Liem, R. K. H. (1986) *J. Neurosci.*, in the press.
30. Powell, B. C., Cam, G. R., Fietz, M. J. & Rogers, G. A. (1986) *Proc. Natl Acad. Sci. USA* 83, 5048–5052.
31. Quax, W., Van der Broeck, L., Egberts, W. V., Ramaekers, F. & Bloemendal, H. (1985) *Cell* 43, 327–338.
32. Quax, W., Egberts, W. V., Hendriks, W., Quax-Jueken, Y. & Bloemendal, H. (1983) *Cell* 35, 215–223.
33. Rigby, P. W. J., Dieckmann, M., Rhodes, C. & Berg, P. (1977) *J. Mol. Biol.* 113, 231–251.
34. Sanger, F., Coulson, A. R., Barrell, B. G., Smith, A. J. H. & Roe, B. (1980) *J. Mol. Biol.* 143, 161–178.
35. Shaw, G., Osborn, M. & Weber, K. (1981) *J. Cell Biol.* 26, 68–82.
36. Shaw, G. & Weber, K. (1981) *Nature (Lond.)* 298, 277–279.
37. Soares, M. B., Scherr, E., Henderson, A., Karathanasis, S. K., Cate, R., Zeitlin, S., Chirgwin, J. & Efstratiadis, A. (1985) *Mol. Cell Biol.* 5, 2090–2103.
38. Southern, E. (1975) *J. Mol. Biol.* 98, 503–517.
39. Tyner, A. L., Eichman, M. J. & Fuchs, E. (1985) *Proc. Natl Acad. Sci.* 82, 4683–4687.
40. Weber, K., Shaw, G., Osborn, M., Debus, E. & Geisler, N. (1983) *Cold Spring Harbor Symp. Quant. Biol.* 48, 717–729.