The structure and organization of the human heavy neurofilament subunit (NF-H) and the gene encoding it

Janice F. Lees¹, Paul S. Shneidman,
Suzanne F. Skuntz, Martin J. Carden²
and Robert A. Lazzarini

The Laboratory of Molecular Genetics, National Institute of
Neurological and Communicative Disorders and Stroke, National
Institutes of Health, Bethesda, MD 20892, USA

¹Present address: Paterson Laboratories, Christie Hospital and Holt
Radium Institute, Wilmslow Road, Manchester M20 9BX, UK

²Guest Researcher from the Department of Pathology and Laboratory
Medicine, University of Pennsylvania, Philadelphia, PA 19104-6079,
USA

Communicated by K. Weber

Genomic clones for the largest human neurofilament protein (NF-H) were isolated, the intron/exon boundaries mapped and the entire protein-coding regions (exons) sequenced. The predicted protein contains a central region that obeys the structural criteria identified for α -helical 'rod' domains typically present in all IF protein components: it is ~310 amino acids long, shares amino acid sequence homology with other IF protein rod domains and displays the characteristic heptad repeats of apolar amino acids which facilitate coiled-coil interaction. Nevertheless, anomalies are noted in the structure of the NF-H rod which could explain observations of its poor homopolymeric assembly *in vitro*. The protein segment on the carboxy-terminal side of the human NF-H rod is uniquely long (>600 amino acids) compared to other IF proteins and is highly charged (>24% Glu, >25% Lys), rich in proline (>12%) and impoverished in cysteine, methionine and aromatic amino acids. Its most remarkable feature is a repetitive sequence that covers more than half its length and includes the sequence motif, Lys-Ser-Pro (KSP) >40 times. Together with the recent identification of the serine in KSP as the main target for NF-directed protein kinases *in vivo*, this repetitive character explains the massive phosphorylation of the NF-H subunit that can occur in axons. The human NF-H gene has three introns, two of which interrupt the protein-coding sequence at identical points to introns in the genes for the two smaller NF proteins, NF-M and NF-L. Both these introns differ from any of the several introns that have a common organization in all other (non-neural) IF genes. However, a clear evolutionary relationship between neural and non-neural IFs is now revealed by the observation here that the third intron in the NF-H gene matches the position of one of the conserved introns in the non-neural IF gene pattern. Hence, divergence of the two IF lineages (neural from non-neural) is more likely to have involved ancestral IF gene duplication rather than RNA-mediated transposition.

Key words: neurofilament/evolution/phosphorylation/
repeated sequence/gene organization

Introduction

Neurofilaments (NFs), the intermediate filaments (IFs) specific to neurons, are the predominant cytoskeletal elements in many nerve cells. They are believed not only to confer mechanical stability, but also to play a role in determining axonal caliber (reviewed by Eagles *et al.*, 1988). Human NFs are composed of three protein subunits, NF-L, NF-M and NF-H, which have mol. wts predicted from their gene sequences of ~62 000, ~102 000 and ~110 000 respectively (Julien *et al.*, 1987; Myers *et al.*, 1987, this report). Structural analyses of the mammalian NF polypeptides (Geisler *et al.*, 1983, 1985a,b) show that they share the same basic architecture as other IF subunit proteins (reviewed by Steinert and Roop, 1988). Each contains a central α -helical 'rod' domain of ~310 amino acids, characterized by a heptad repeat in the amino acid sequence in which the first and fourth residues are apolar. Such a molecular arrangement allows the formation of 10-nm filaments through a coiled-coil interaction of the α -helices (Crick, 1953; Cohen and Parry, 1986). While the rod domains of IF proteins are structurally conserved, the amino- and carboxy-terminal sequences which flank them have diverged dramatically. The uniqueness of these amino- and carboxy-terminal domains supports the idea rooted in biochemical studies that they confer functional specificity on the IFs. Indeed, a prominent feature of both NF-M and NF-H is a highly charged, heavily phosphorylated (9–100 mol phosphate/mol protein) carboxy-terminal extension (Chin *et al.*, 1983; Geisler *et al.*, 1983; Julien and Mushynski, 1983; Carden *et al.*, 1985) not shared by other IFs. This domain has been proposed as part of the structural component of the intermediate cross-bridges observed in axons (Hirokawa *et al.*, 1984). There is an intriguing correlation in this respect between the location of NFs within neurons and the degree to which they are phosphorylated. Perikaryal NFs are predominantly unphosphorylated, while axonal NFs are highly phosphorylated (Sternberger and Sternberger, 1983; Bennett and DiLullo, 1985; Lee *et al.*, 1986, 1987; Carden *et al.*, 1987). However, the precise role of NF phosphorylation remains to be elucidated.

Recently, we reported the sequence of the human NF-M subunit predicted from cDNA clones (Myers *et al.*, 1987). A striking feature of its sequence was a 13 amino acid sequence tandemly repeated six times. The speculation that the 12 serine residues contained in this repeated sequence were the sites of phosphorylation has been borne out by subsequent chemical analysis of tryptic fragments of NF-M (Geisler *et al.*, 1987) and by immunologic analysis employing synthetic oligopeptides composed of the repeated sequence (Lee *et al.*, 1988).

The structures of both human and mouse NF-M and NF-L genes have been determined (Lewis and Cowan, 1985; Julien *et al.*, 1987; Levy *et al.*, 1987; Myers *et al.*, 1987) and shown to differ significantly from the organization of

the genes encoding all other (non-neuronal) IF subunits. While the striking homology in amino acid sequence and structure of all IF proteins strongly suggests that their genes were derived from a common ancestor, the evolutionary pathway giving rise to the NF gene group is obscure and controversial. To account for the limited number of introns observed in NF-L as compared to the other IF genes, Lewis and Cowan (1986) proposed that an RNA-mediated transposition event took place early in the evolution of the NF genes. An alternative interpretation, offered by Myers *et al.* (1987), was based largely on their analysis of the NF-M gene. They favor the notion that the divergence of the neuronal intermediate filaments (and perhaps lamins) from the ancestral IF gene was a very early event which occurred prior to the divergence of keratins, vimentin, desmin and glial fibrillary acidic protein. The presence of NFs in the earliest Metazoa can be assumed since they are present in both contemporary vertebrate and invertebrate species, with the possible exception of the arthropods (Lasek *et al.*, 1985). From this perspective, the differences in gene organization between NF proteins and other contemporary IF subunits reflect the very large evolutionary distance between the present and the time of divergence.

Unfortunately, an evaluation of the various proposals for the evolution of the NF genes is severely hampered by the limited number of IF genes whose organizations have been studied. In particular, the absence of detailed information on the genomic organization of the gene encoding the largest NF subunit (NF-H) has been particularly restrictive. We have isolated genomic clones from human NF-H and report, for the first time, the nucleotide sequence of its complete coding region together with an analysis of the gene structure. The data reveal a protein with highly unusual properties, the most striking of which is a highly repetitive region spanning more than one half of the carboxy-terminal extension and containing >40 potential phosphorylation sites. The location of introns in the NF-H gene makes it unlikely that a unique RNA-mediated transposition event occurred after the divergence of the NFs from the ancestral IF gene and accounts for the differences in the organization of the NF genes.

Results

Isolation and characterization of genomic clones encoding human NF-H

Initial screening of a human liver (λ EMBL3) genomic library for NF-H-specific clones was performed using a 104-bp *PvuII*–*RsaI* restriction fragment from exon 2 of the human NF-M gene (Myers *et al.*, 1987). This probe was selected because it encodes the most highly conserved portion of the rod domain, one common to all IF protein subunits. Hybridization with this probe at reduced stringencies identified two clones, HW10 and HW12, which were shown by cross-hybridization studies to be distinct from either NF-L or NF-M gene. An additional clone, HW14, was subsequently isolated using a cDNA clone of mouse NF-H mRNA (P.S.Shneidman *et al.*, in preparation). Restriction endonuclease maps were generated from all three genomic clones and established that they overlapped as shown in Figure 1. Appropriate restriction fragments were subcloned into

Fig. 1. Identification of human NF-H genomic clones and restriction map of the entire gene. The map above, which is drawn to scale, was constructed from digestion studies using the restriction endonucleases *Bam*HI (B), *Hind*III (H), *Kpn*I (K), *Sac*I (S) and *Xba*I (X). Three genomic clones for human NF-H, HW10, HW12 and HW14 were identified by reduced-stringency hybridization with an oligonucleotide probe based on exon 2 of human NF-M (Myers *et al.*, 1987), corresponding to exon 3 of the NF-H gene as indicated by the dotted line, or with a partial mouse NF-H cDNA spanning all three introns (P.S.Shneidman *et al.*, in preparation). The two asterisks indicate an ambiguous assignment of a single *Bam*HI site. Introns were not sequenced, their sizes being inferred from fragment migration in agarose gels: intron 1, ~2.6 kb; intron 2, ~2.0 kb; intron 3, ~4.0 kb (all sizes $\pm$ 10%).

Fig. 2. Northern blot analysis of human spinal cord RNA probed for NF-H and NF-M mRNAs. Numbers indicate the sizes (kb), of the RNA standards (left lane) and the NF mRNAs (centre and right-hand lanes). NF-M was detected using a cDNA (pHNF11) encoding the second half of the rod plus the initial 150 residues of the carboxy-terminal domain from human NF-H (see Myers *et al.*, 1987). For NF-H, the probe was an 1150 bp *Pst*I–*Pst*I fragment of clone HW12 (Figure 1) representing the 5' half of exon 4: the first *Pst*I site is at the end of the third intron, within 10–20 bp of the splice junction (data not shown), while the second *Pst*I recognition site (CTGCAG) can be seen at nucleotides 2330–2335 in Figure 3.

pTZ18/19 or M13mp18/19 for sequencing and further analysis.

The major species of mRNA derived from the NF-H gene is slightly larger than the mRNA encoded by the NF-M gene. Figure 2 presents a Northern transfer analysis of total RNA isolated from spinal cord probed with either a restriction endonuclease fragment derived from HW12 (NF-H) or pNF11 a probe for human NF-M (Myers *et al.*, 1987). This

analysis indicates that the NF-H mRNA is ~3.9 kb in length while that of NF-M is ~3.5 kb.

Nucleotide sequencing primarily of HW10 and HW12 was performed using the dideoxy chain termination method (Sanger *et al.*, 1980) and synthetic oligonucleotides complementary to the appropriate regions of the cloned DNA. The sequence of both DNA strands were independently determined for each exon and for short parts of the flanking intron regions (20–100 nt in each case). The human NF-H gene is composed of four exons and restriction analysis shows that it spans ~12 kb (Figure 1). Since NF-H has an additional intron beyond those found in NF-M, exon 3 of NF-H is homologous to exon 2 of NF-M and was originally identified by hybridization with probes derived from exon 2 of NF-M. Intron positions were located using mouse NF-H cDNAs spanning from the middle of exon 1 to the end of the gene (P.S.Shneidman *et al.*, in preparation) and by comparing predicted protein sequence with published partial sequence of NF-H from pig (Geisler *et al.*, 1985a) and rat (Robinson *et al.*, 1986). A synthetic oligonucleotide based on sequence from the 5' end of a mouse NF-H cDNA clone, pMuH1 (P.S.Shneidman *et al.*, in preparation), was used as a probe to identify the first exon in restriction digests of human NF-H genomic clones and to sequence the resulting fragments.

Comparison of the human NF-L (Julien *et al.*, 1987) and NF-M (Myers *et al.*, 1987) amino acid sequences with the one encoded by clones HW10 and HW14 suggests that NF-H translation begins at position 1 or 2 in Figure 3. This reading frame is closed by two termination codons further upstream (positions -72 and -93) and the methionines at positions 1 and 2 are the first in the only one reading frame large enough to encode NF-H. Comparison of the sequence in this region with that surrounding the human NF-M initiating codon ((Myers *et al.*, 1987) implicates the second, not the first, ATG as the actual initiating codon for NF-H. This assignment is supported by the studies of Kozak (1987) which indicate that the second ATG is presented in a preferred sequence context. A stop codon (TAA) located at position 3342 after the first ATG predicts a protein of 1019 (or 1018) residues having a mol. wt of ~110 000. A polyadenylation consensus sequence is located at position (3620) ~20 bp upstream from the 3' end of the mRNA as determined from a mouse NF-H cDNA clone (P.S.Shneidman *et al.*, in preparation).

Predicted protein sequence

Like all other IF subunit polypeptides, the predicted sequence of human NF-H can be subdivided into three regions: an amino-terminal domain of ~100 amino acids, a central helical rod region of ~310 amino acids and a carboxy-terminal domain.

The amino-terminal domain is particularly rich in serine (19%), glycine (16%) and alanine (15%), and provides clear evidence that the clones HW10 and HW12 correspond to the human NF-H gene: amino acids 31–80 in the predicted sequence (Figure 3) correspond with >95% homology to tryptic peptides T₂, T₄, T₃, T₁ and T₅ respectively derived by direct protein sequencing of an amino-terminal fragment from porcine NF-H (Geisler *et al.*, 1985a).

By analogy with other IFs (Geisler and Weber, 1985; Steinert and Roop, 1988), a central 'rod' domain can be

delineated in the predicted human NF-H sequence extending from amino acid 101 to 410. This rod region is characterized by a quasi-heptad repeat (*a-b-c-d-e-f-g*) in which positions *a* and *d* are generally occupied by apolar residues as illustrated by the dots in Figure 3. Assignment of these positions was made in accordance with previous analyses (e.g. Geisler *et al.*, 1985a), although it should be emphasized that certain features of the NF-H rod distinguish it from the rods of other IF proteins. Like that of NF-M (Myers *et al.*, 1987), lamins (Fisher *et al.*, 1986) and certain keratins (e.g. Franz and Franke, 1986; Miyatani *et al.*, 1986), the NF-H rod contains no proline residues. Prolines, when they are present in other IF protein rods, define 'linker', non-helical segments that have been proposed as interrupting otherwise conserved α -helical subsegments of the rod region (Steinert and Parry, 1985). Thus, linker L₁ appears absent from NF-H, as observed previously for NF-M (Myers *et al.*, 1987), so that the first helix segment may be longer in NF-M and NF-H than in other IF proteins. A second peculiarity of the NF-H rod is its richness in charged amino acids or those with only small apolar side chains between positions 186 and 215. Consequently, it is not easy to assign a heptad pattern consistent with coiled-coil forming ability (Cohen and Parry, 1986) within this part of the rod. A third unusual feature of the NF-H rod is its acquisition of five extra amino acids (GSGAA—position 250 compared to NF-M and NF-L). Despite the discrepancies just described, the distal part of the rod domain corresponding to helix 2b obeys most of the structural rules delineated for all other IF subunits including heptad repeats, length and a change in the phase of the heptad repeat (Steinert and Roop, 1988).

Immediately following the YRKLLGEE (residues 402–410; Figure 3) sequence defining the end of the rod, the NF-H sequence diverges dramatically from those of other IF subunits, including NF-M and NF-L. A short (~30 amino acid) segment containing five prolines begins the carboxy-terminal domain. This initial segment contains the major site for chymotryptic cleavage (Geisler *et al.*, 1985) which liberates the entire carboxy-terminal domain of NF-H (Chin *et al.*, 1983; Julien and Mushynski, 1983; Carden *et al.*, 1985). Proteolytic sensitivity, proline content and bulky, hydrophobic side chains characterize this segment of the NF-H protein. Following it is a highly charged region (~60 amino acids) containing successive tracts of glutamic acid (E) residues. An analogous sequence is found at the corresponding positions in NF-M and NF-L. The most remarkable segment of the carboxy-terminal domain, however immediately follows this Glu-rich region and consists of slightly degenerate repeats based on the sequence EAKSPEKAKSPVKE. Indeed, serine (S) followed by proline (P), and generally preceded by lysine (K) occurs 41 times. Figure 4 shows these repeats in detail. A central area of direct homology is flanked on both sides by regions where the repeat structure is degenerate, becoming more so towards each end. Following the repeat region is a segment ~200 amino acids long that continues to the carboxy-terminus of NF-H. This last region of the NF-H carboxy-terminal domain maintains the composition of the repeat region with one important exception, i.e. a dearth of serine residues (4.6% compared to 12.4% in the repeat). A very high content of lysine (>30%) makes this last segment slightly basic (Glu = 22%; Asp = 5.5%), despite the total absence of arginine.

1950

```

1962      1977      1992      2007      2022      2037
GAG AAG GCC AAG TCC CCA GAG AAG GAA GAG GCC AAG TCC CCT GAG AAG GCC AAG TCC CCA GTG AAG GCA GAA GCA AAG TCC CCT GAG AAG GCC AAG TCC 2046
E K A K S P E K E E A K S P E K A K S P V K A E A K S P E K A K S 682
-----23-----      -----24-----      -----25-----      -----26-----      -----27-----
2061      2076      2091      2106      2121      2136
CCA GTG AAG GCA GAA GCA AAG TCC CCT GAG AAG GCC AAG TCC CCA GTG AAG GAA GAA GCA AAG TCC CCT GAG AAG GCC AAG TCC 2145
P V K A E A K S P E K A K S P V K E E A K S P E K A K S P V K E E 715
-----28-----      -----29-----      -----30-----      -----31-----
2160      2175      2190      2205      2220      2235
GCA AAG TCC CCT GAG AAG GCC AAG TCC CCA GTG AAG GAA GAA GCA AAG ACC CCC GAG AAG GCC AAG TCC CCA GTG AAG GAA GAA GCC AAG TCC CCA GAG 2244
A K S P E K A K S P V K E E A K T P E A K A K S P V K E E A K S P E 748
-----32-----      -----33-----      -----34-----      -----34-----      -----35-----
2259      2274      2289      2304      2319      2334
AAG GCC AAG TCC CCA GAG AAG GCC AAG ACT CTT GAT GTG AAG TCT CCA GAA GCC AAG ACT CCA GCG AAG GAG GAA GCA AGG TCC CCT GCA GAC AAA TTC 2343
K A K S P E K A K T L D V K S P E A K T P A K E E A R S P A D K F 781
-----36-----      -----37-----      -----38-----      -----39-----
2358      2373      2388      2403      2418      2433
CCT GAA AAG GCC AAA AGC CCT GTC AAG GAG GAG GTC AAG TCC CCA GAG AAG GCG AAA TCT CCC CTG AAG GCG GAT GCC AAG GCC CCT GAG AAG GAG ATC 2442
P E K A K S P V K E E V K S P E K A K S P L K A D A K A P E K E I 814
-----40-----      -----41-----      -----42-----
2457      2472      2487      2502      2517      2532
CCA AAA AAG GAA GAG GTG AAG TCC CCA GTG AAG GAG GAG GAG AAG CCC CAG GAG GTG AAA GTC AAA GAG CCC CCA AAG AAG GCA GAG GAA GAG AAA GCC 2541
P K K E E V K S P V K E E E K P Q E V K V K E P P K K A E E E K A 847
-----43-----
2556      2571      2586      2601      2616      2631
CCT GCC ACA CCA AAA ACA GAG GAG AAG AAG GAC AGC AAG AAA AAG GAG GCA CCC AAG AAG GAG GCT CCA AAG CCC AAG GTG GAG GAG AAG GAA CCT 2640
P A T P K T E E K K D S K K E E A P K K E A P K P K V E E K K E P 880
2655      2670      2685      2700      2715      2730
GCT GTC GAA AAG CCC AAA GAA TCC AAA GTT GAA GCC AAG AAG GAA GAG GCT GAA GAT AAG AAA AAA GTC CCC ACC CCA GAG AAG GAG GCT CCT GCC AAG 2739
A V E K P K E S K V E A K K E E A E D K K K V P T P E K E A P A K 913
2754      2769      2784      2799      2814      2829
GTG GAG GTG AAG GAA GAC GCT AAA CCC AAA GAA AAG ACA GAG GTG GCC AAG AAG GAA CCA GAT GAT GCC AAG GCC AAG GAA CCC AGC AAA CCA GCA GAG 2838
V E V K E D A K P K E K T E V A K K E P D D A K A K E P S K P A E 946
2853      2868      2883      2898      2913      2928
AAG AAG GAG GCA GCA CCG GAG AAA AAA GAC ACC AAG GAG GAG AAG GCC AAG AAG CCT GAG GAG AAA CCC AAG ACA GAG GCC AAA GCC AAG GAA GAT GAC 2937
K K E A A P E K K D T K E E K A K K P E E K P K T E A K A K E D D 979
2952      2967      2982      2997      3012      3027
AAG ACC CTC TCA AAA GAG CCT AGC AAG CCT AAG GCA GAA AAG GCT GAA AAA TCC TCC AGC ACA GAC CAA AAA GAC AGC AAG CCT CCA GAG AAG GCC ACA T 3036
K T L S K E P S K P K A E K A E K S S S T D Q K D S K P P E K A T 1012
3051      3080      3090      3100      3110      3120      3130      3140      3150
GAA GAC AAG GCC GCC AAG GGG AAG TAA GGCAGGG AGAAAGGAAC ATCCGGAACA GCCAAAGAAA CTCAGAAGAG TCCCGGAGCT CAAGGATCAG AGTAACACAA TTTCACTTT 3150
E D K A A K G K * 1020
TTCTGCTTT ATGTAAGAAG AAATGCTTA GATGACGGGG CCTCTCTCTT CAAACAGGAA TTCTGTGTAG CAATATGTGA GCAAGAGAGG GCACTCCAG GCCCTGCC CCAGGCCCTC 3270
CCCAGGCGAT GGACAATTAT GATAGCTTAT GTAGCTGAAT GTGATACATG CCGAATGCCA CACGTAACAA CTGACTATA AAAAATGCC CCTCTCTTC CAAATAAGTG CATTATTGCG 3390
CTCTATGTGC AACTGACAGA TGACCGCAAT AATGAATGAG CAGTTAGAAA TACATTATGC TTGAGATGTC TTAACCTATT CCCAAATGCC TTCTGTTTC CAAAGGAGTG GTCAAGCCCT 3510
TGCCAGAGC TCTCTATTCT GGAAGAGCGG TCCAGGTGGG GCCGGGACT GGCCACTGAA TTATGCCAGG GCGCACTTTC CACTGGAGTT CACTTTCAAT TGCTTCTGTG CAATAAACC 3630
AAGTGCTTAT AAAATGAAAA TGTGTCTGCT GTTATTCTCT TTCCCTGGGA AGGCTGGGGG CAGGGCAGGG GAGGTCTGGA TGTGACACC CAGACTGCAT GGGACTGAGC AAGCATCAGT 3750

```

Fig. 3. Nucleotide sequence of the coding regions in the human NF-H gene and the predicted protein sequence. Translation of the longest open reading frame is presented in single letter code. Filled circles indicate apolar amino acids in the heptad repeat of the rod domain; open circles indicate amino acids at variance with the heptad repeat theme. Each element of the carboxy-terminal domain multi-phosphorylation site is numbered and underscored with a dashed line.

of two of the NF-H introns correspond precisely to the positions of the introns in the other human neurofilament subunits NF-L and NF-M. The position of the remaining NF-H intron is unique among the neurofilaments, but it is very close (within one amino acid) to the position of an intron found in all other intermediate filament genes examined thus far. The relationship between the intron positions and the structural domains of the IFs is given in the upper panel of Figure 5. In this figure, scaled graphic representations of several IFs are positioned so that homologous sequences of the α -helical domain are aligned vertically. The carats indicate the points at which the coding sequence is interrupted by introns. It can be readily appreciated from this figure that the genes encoding the three NF subunits are closely related to one another. Furthermore, the relationship of NF-H to the other IFs is apparent—the position of the first and third NF-H introns line up close to the intron positions found in the other mammalian IFs. A detailed comparison of these two sites of intron positions is shown in the lower panels of Figure 5. The left panel illustrates that the first NF-H intron interrupts the coding sequence at a position only one amino acid away from the intron position found in all IFs except NF-L and NF-M. This misalignment is well within

the range of differences observed between the intron positions of undisputedly related genes. For example, the same panel of Figure 5 shows that the corresponding intron positions in keratins I and II differ by three amino acids. The alignment of the third intron on NF-H with introns of the other IFs is less precise, but still suggestive of homology in gene organization.

Finally, it should be noted that the NF-L gene carries an intron that interrupts the glutamate-rich coding region near the carboxy terminus. This interruption is unique and not shared by the genes encoding the two larger neurofilament subunits. The latter, however, carry the repetitive phosphorylation sites that interrupt the glutamate-rich regions. These observations suggest the possibility that the sequence encoding the tandem repeat of NF-M and NF-H may have evolved from an intron that interrupted the region in an ancestral NF-L like gene. A second attractive scenario for the origin of the tandem repeat is that it arose not from the intron, but from sequences encoding the glutamate-rich domain by mutation and reiteration. The fact that the tandem repeat is rich in glutamate and that the glutamate residues most frequently occur as the di- or tripeptides (E)₂ or (E)₃ in all three regions (the glutamate-rich, the tandem repeat

K	S	P	E	K	A	K	S	P	V	K	E	E	A	
				>>>>>>>>	K	S	P	P	A	E	E	A		
A	S	P	E	K	E	A	K	S	P	V	K	E	E	A
K	S	P	A	E	A	K	S	P	E	K	E	E	A	A
K	S	P	A	E	V	-	-	-	-	-	-	-	-	-
K	S	P	E	K	A	K	S	P	A	K	E	E	A	A
K	S	P	P	E	A	K	S	P	E	K	E	E	A	A
K	S	P	A	E	V	-	-	-	-	-	-	-	-	-
K	S	P	E	K	A	K	S	P	A	K	E	E	A	A
K	S	P	A	E	A	-	-	-	-	-	-	-	-	-
K	S	P	E	K	A	K	S	P	V	K	E	E	A	A
K	S	P	A	E	A	K	S	P	V	K	E	E	A	A
K	S	P	A	E	V	-	-	-	-	-	-	-	-	-
K	S	P	E	K	A	K	S	P	T	K	E	E	A	A
K	S	P	E	K	A	K	S	P	E	K	E	E	A	A
K	S	P	E	K	A	K	S	P	V	K	A	E	A	A
K	S	P	E	K	A	K	S	P	V	K	A	E	A	A
K	S	P	E	K	A	K	S	P	V	K	E	E	A	A
K	S	P	E	K	A	K	S	P	V	K	E	E	A	A
K	T	P	E	K	A	K	S	P	V	K	E	E	A	A
K	S	P	E	K	A	-	-	-	-	-	-	-	-	-
K	S	P	E	K	A	-	-	-	-	-	-	-	-	-
K	T	L	D	-	V	-	-	-	-	-	-	-	-	-
K	S	P	E	-	A	K	T	P	A	K	E	E	A	A
R	S	P	A	D	-	-	-	-	-	-	-	-	-	-
K	F	P	E	K	A	K	S	P	V	K	E	E	V	
K	S	P	E	K	A	K	S	P	L	K	A	D	A	
K	A	P	E	K	-	E	I	P	K	K	E	E	V	
-	-	-	-	-	-	K	S	P	V	K	E	E	>>>>>>	

Fig. 4. Predicted protein sequence of the repeat (phosphorylation) region in the carboxy-terminal domain of human NF-H.

and the end region) of both the NF-H and NF-M (Myers *et al.*, 1987) carboxy-terminal domains is consistent with this second interpretation. In the absence of additional information, we cannot favor one suggestion over the other.

Discussion

This report describes for the first time the entire protein sequence of human NF-H, information long awaited to explain certain properties of the largest NF subunit revealed by detailed biochemical, immunological and cell biological analyses. These include its inability to self-assemble *in vitro*, the large size and highly charged nature of its carboxy-terminal (peripheral) domain and, above all, its remarkably extensive phosphorylation.

We confirm that NF-H indeed possesses a central rod domain (Geisler *et al.*, 1983) and show that it has the overall properties attributed to all other IF protein subunits. Hence, NF-H appears able to co-polymerize with NF subunits as an integral part of filaments rather than to attach by a looser association. Nevertheless, structural anomalies of the NF-H rod domain sequence compared to other IF proteins (noted in the results) suggest that it might not be as well suited to the formation of coiled-coil interactions (see also Geisler *et al.*, 1985a). These anomalies in the NF-H rod might explain both its poor reassembly *in vitro* (Geisler and Weber, 1981) and the fact that it is a monomer at low ionic strength (Cohlberg *et al.*, 1987), while desmin and vimentin are tetramers or higher oligomers under the same conditions (Geisler and Weber, 1985). Because of the potential weakness in coiled-coil forming capacity of the NF-H, the implications are that NF-H might be an obligate heteropolymeric IF subunit like keratins (Geisler and Weber, 1985; Steinert and Roop, 1988), an observation borne out by its developmental profile in neurons, appearing as it does only

after prior expression of NF-L and NF-M (Carden *et al.*, 1987 and references therein). Nevertheless, the potential effect of NF-H rod domain structure upon its ability to assemble as a homopolymer intermediate-sized filament must await further study, such as in transfection studies of the gene into cells lacking NF proteins, or after *in vitro* expression of the nascent (i.e. unmodified by phosphorylation), non-denatured protein.

The unusual character of the NF-H carboxy-terminal domain is also now revealed to be extremely highly charged, containing 24.1% glutamic acid and 25.7% lysine, and punctuated by many proline residues. This composition suggests an extended structure that is highly hydrated. This suggestion is corroborated by the ease with which NF forms viscous gels (Carden and Eagles, 1983; Leterrier and Eyer, 1987). Apart from its first 30 amino acids, the NF-H carboxy-terminal domain is strikingly impoverished in cysteine, methionine, aromatic residues and arginine residues. This amino acid composition explains the NF-H resistance to many proteases and cleavage reagents commonly used for protein chemical analyses (Geisler *et al.*, 1985). Its high Glu/Lys content explains its high susceptibility to V8 protease and endo Lys-C (M.J. Carden and W.W. Schlaepfer, unpublished observations). The unusual, highly charged composition of the peripheral (carboxy-terminal) domain also explains the abnormal migration of NF-H in SDS gels which leads to gross overestimates of its mol. wt (200–220 000) that are accentuated, but not wholly explained, by the substantial phosphorylation of this protein *in vivo* (e.g. Georges and Mushynski, 1987). The mol. wt of 110–114 000 (depending upon the extent of phosphorylation) predicted here agrees well with values obtained by gel filtration and sedimentation (Kaufmann *et al.*, 1984) and by SDS-PAGE after chemical neutralization of charge and enzymatic dephosphorylation (Georges and Mushynski, 1987).

Probably the most interesting feature of the NF-H carboxy-terminal domain, however, is its repeat region. Convincing evidence now exists for the idea that the serine residues in this repeat are the major sites for *in vivo* phosphorylation of the protein (Geisler *et al.*, 1987; Lee *et al.*, 1988). Hence, precise sequence information will greatly facilitate the identification of the specific kinases that phosphorylate NFs *in vivo*. An examination of the sequences of NF-M carboxy-terminal domains from human (Myers *et al.*, 1987), chicken (Zopf *et al.*, 1987), rat (Napolitano *et al.*, 1987) and mouse (Levy *et al.*, 1987) proteins reveals that some species have evolved repeat regions in this mid-sized NF protein while others have not. The exact sequence of these repeat regions differs markedly between species, but the flanking sequences are quite tightly conserved. Such variations of the repeat theme suggest that they are hypervariable. That this is also true for the repeat in NF-H is indicated by a comparison of the human, rat (Robinson *et al.*, 1987), mouse (P.S. Shneidman *et al.*, in preparation) and pig (Geisler *et al.*, 1985a, 1987) NF-H sequences. This variability has important implications for both the differential function of NFs among species and also for revealing the mechanism of how carboxy-terminal NF domains are evolving. Certainly, immunological cross-reactivity suggests that most species have developed a multiphosphorylation (repeat) region in the carboxy-terminal extension of their largest NF subunit (Lee *et al.*, 1988). The fact that this repeat is also

Fig. 5. A comparison of intron positions in the three human NF genes compared to the genes encoding other (non-neural) IF subunits. Sequence information was obtained from Julien *et al.* (1987) (NF-L); Myers *et al.* (1987) (NF-M); this report (NF-H); Ferrari *et al.* (1986) (Vim); Balcarek and Cowan (1985) (GFAP); Quax *et al.* (1985) (Des); Marchuk *et al.* (1985) (Ker 1); and Johnson *et al.* (1985) (Ker 2).

displayed by the NF-M subunit in some species, but not in others, suggests either that it arose independently in some species (convergent evolution) or that this region is not only hypervariable, but subject to deletion or insertion.

The NF subunits bear striking similarities to one another, strongly indicating that they arose from a common progenitor neuronal IF gene by gene duplication. Each human NF gene has two introns located precisely in the same position relative to the coding sequence (Figure 5), suggesting that the progenitor gene contained at least these two introns prior to the duplication. The alternative, that these introns were introduced after duplication and distributed among the three subunit genes by gene conversion, seems unlikely. While homology between the NF proteins is apparent, the nucleotide sequence homology is not high. Proposed mechanisms of gene conversion generally involve strand displacement, branch migration and resolution between nearly identical genes resulting ultimately in the formation of two completely homologous genes (Slightom *et al.*, 1980; Smithies and Powers, 1986). Since, in the present case, such conversion would require two events involving three genes mapping to two different chromosomes (D.Nelson *et al.*, unpublished observations), we discount this possibility. That NF-L and NF-H genes each carry additional introns beyond the two common to all NF subunit genes suggest that the progenitor gene may have contained more than just two introns and that some of these were lost after the gene duplication. Alternatively, the additional introns could have been added after duplication by as yet unknown mechanisms. At present, information about the mechanisms of intron entry and departure from genes is scant, but evidence that these

processes occur is plentiful. This evidence is most easily gleaned from the comparison of the organization of genes that arose from common ancestors such as the insulin genes (Perler *et al.*, 1980; Soares *et al.*, 1985), leghemoglobin and myoglobin (Wiborg *et al.*, 1982), α and β actin (Fornwald *et al.*, 1982; Zakut *et al.*, 1982), α and β tubulin (reviewed by Cleveland and Sullivan, 1985), and keratins I and II (Fuchs and Marchuk, 1983; Lehnert *et al.*, 1984; Steinert and Roop, 1988). In each of these families of genes some members contain introns not found in the others. The neurofilament family is clearly yet another example of this phenomenon. Often the intron positions in related genes are misaligned slightly, suggesting that the intron migrated. Comparisons of the keratin II genes to other IFs (Figure 5) or the genes encoding chicken and mouse carbonic anhydrase (Yashihara *et al.*, 1987), insulin, IGF I and IGF II (Soares *et al.*, 1986) reveal examples of intron 'sliding' or 'evolutionary adjustments' in intron positions. The IF genes also offer examples of this type of intron migration as well. Intron 1 of the NF-H is positioned within one amino acid of the position found in other IF genes. Similarly, one of the introns of the keratin genes is also slightly displaced from the position used in other IFs (Figure 5). The other intron positions are more difficult to evaluate. The NF intron positions shown in the right panel of Figure 5 may be homologous to the position of intron 6 in desmin, GFAP and vimentin, but the displacement is eight amino acids, considerably larger than most of the examples of intron sliding. A final decision as to the relatedness of these positions may have to await more detailed knowledge of the mechanisms of intron sliding.

The differences between the organization of the NF and other IF genes are obvious—NF genes have two or three introns, other IF genes have seven or eight. Lewis and Cowan (1985) and Levy *et al.* (1987) have suggested that transposition of mRNA sequences back to the genome via a retroposon may account for the differences in gene organization. The work presented here and earlier (Myers *et al.*, 1987) allow us to place certain limitations on the time and extent of the proposed RNA transposition. Since all three NF subunit genes have an homologous organization, any RNA transposition event is likely to have occurred before the gene was duplicated. Since NF-H has retained an intron in a position homologous to one found in all other IFs, it seems likely that only a portion of the present day NF gene could have originated from an mRNA-mediated transposition. In the absence of the two hallmarks of RNA transposition (a 3' remnant of a polyadenylated tail and direct repeat sequences flanking the gene in question), we conclude that it is unlikely that a simple RNA transposition accounts for the reduced number of introns in neurofilament genes compared to other IF genes.

Materials and methods

Isolation and analysis of genomic clones

NF-H clones were isolated from a human genomic library prepared in λ EMBL3 (generously supplied by J. Weiss, see Frischauf, 1987). Nylon filters were prepared from 150-mm plates each containing 3×10^4 plaques and probed with a 104-bp *PvuII*–*RsaI* restriction fragment derived from the exon 2 region of human NF-M (Myers *et al.*, 1987). Hybridization was performed at reduced stringency in 40% formamide, $5 \times$ SSC, 1% SDS, 50 mM sodium pyrophosphate, $5 \times$ Denhardt's solution, 100 mg/ml salmon sperm DNA at 42°C for 18 h followed by washing in $0.1 \times$ SSC, 0.1% SDS at 55°C (Maniatis *et al.*, 1982).

Restriction maps of DNA inserts were initially constructed using the partial digestion technique (Rackwitz *et al.*, 1984) and confirmed using appropriate double digests. Restriction fragment were subcloned into pTZ18/19 (Pharmacia) or M13mp18/19 vectors for subsequent analyses. The nucleotide RNA sequence was determined using the dideoxy chain termination method (Sanger *et al.*, 1980) employing both the universal primer and synthetic oligonucleotides.

RNA analysis

RNA was isolated from human spinal cord using the guanidinium isothiocyanate method as described by Maniatis *et al.* (1982). RNA was denatured in formaldehyde and electrophoresed in 1% agarose gels containing 0.04 M MOPS, pH 7.0, 2.2 M formaldehyde (Maniatis *et al.*, 1982). [14 C]rRNA or [14 C]tRNA transcripts of known size (Bethesda Research Labs) were used as molecular length standards. Fractionated RNA was transferred to nylon membranes by capillary blotting. The membrane transfers were hybridized under the conditions of high stringency given above, using the indicated 32 P-labeled probes. Membranes were washed as above and autoradiographs prepared.

Acknowledgements

We wish to thank Jean-Pierre Julien for making his clone of the third exon of the mouse NF-H gene available to use before its formal published description, John Weiss for the generous contribution of the human genomic library, John Trojanowski for human brain tissue, and David Nelson, Bill Schlaepfer, Virginia Lee and Klaus Weber for their valuable advice and discussion. We would like to thank Charlene French and Judy Carrillo for their assistance in the preparation of the manuscript. Martin Carden was supported by NIH grant PO-AG06107 (P.I. Bill Schlaepfer).

References

- 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. and Eagles, P.A.M. (1983) *Biochem. J.*, **215**, 227–237.
- Carden, M.J., Schlaepfer, W.W. and Lee, V.M.-Y. (1985) *J. Biol. Chem.*, **260**, 9805–9817.
- Carden, M.J., Trojanowski, J.Q., Schlaepfer, W.W. and Lee, V.M.-Y. (1987) *J. Neurosci.*, **7**, 3489–3501.
- Chin, T.-K., Eagles, P.A.M. and Maggs, A. (1983) *Biochem. J.*, **215**, 239–252.
- Cleveland, D.W. and Sullivan, K.F. (1985) *Annu. Rev. Biochem.*, **65**, 331–365.
- Cohen, C. and Parry, D.A.D. (1986) *Trends Biochem. Sci.*, **11**, 245–248.
- Cohlberg, J.A., Hajarian, H. and Sainte-Marie, S. (1987) *J. Biol. Chem.*, **262**, 17009–17015.
- Crick, F.H.C. (1953) *Acta Crystallogr.*, **6**, 685–688.
- Eagles, P.A.M., Pant, H.C. and Gainer, H. (1988) In Goldman, R.D. and Steinert, P.M. (eds), *Intermediate Filaments*. Plenum Press, New York, in press.
- Ferrari, S., Battini, R., Kaczmarek, L., Rittling, S., Calabretta, B., Kim de Riel, J., Philiponis, V., Wei, J.-F. and Baserga, R. (1986) *Mol. Cell Biol.*, **6**, 3614–3620.
- Fisher, D.Z., Chaudhary, N. and Blobel, G. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 6450–6454.
- Fornwald, J.A., Kuncio, G., Peng, I. and Ordahl, C.P. (1982) *Nucleic Acids Res.*, **10**, 3861–3876.
- Franz, J.K. and Franke, W.W. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 6475–6479.
- Frischauf, A.M. (1987) *Methods Enzymol.*, **152**, 190–199.
- Fuchs, E. and Marchuk, D. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 5857–5861.
- Geisler, N. and Weber, K. (1985) In Shay, J.W. (ed.), *Cell and Molecular Biology of the Cytoskeleton*. Plenum Press, New York, pp. 41–68.
- Geisler, N., Kaufmann, E., Fischer, S., Plessmann, U. and Weber, K. (1983) *EMBO J.*, **2**, 1295–1302.
- 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.
- Geisler, N., Vandekerckhove, J. and Weber, K. (1987) *FEBS Lett.*, **221**, 403–407.
- Georges, E. and Mushynski, W.E. (1987) *Eur. J. Biochem.*, **165**, 281–287.
- Hirokawa, N., Glicksman, M.A. and Willard, M.B. (1984) *J. Cell Biol.*, **98**, 1523–1536.
- Johnson, L.D., Idler, W.W., Zhou, X.-M., Roop, D.R. and Steinert, P. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 1896–1900.
- Julien, J.-P. and Mushynski, W.E. (1983) *J. Biol. Chem.*, **258**, 4019–4025.
- Julien, J.-P., Meyer, D., Flavell, D., Hurst, J. and Grosveld, F. (1986) *Mol. Brain Res.*, **1**, 243–250.
- Julien, J.-P., Grosveld, F., Yazdanbakhsh, K., Flavell, D., Meiser, D. and Mushynski, W.E. (1987) *Biochim. Biophys. Acta*, **909**, 10–20.
- Kaufmann, E., Geisler, N. and Weber, K. (1984) *FEBS Lett.*, **170**, 81–84.
- Kozak, M. (1987) *Nucleic Acids Res.*, **15**, 8125–8147.
- Lasek, R.J., Phillips, L., Katz, M.J. and Autilio-Gambetti, L. (1985) *Ann. NY Acad. Sci.*, **455**, 462–478.
- Lee, V.M.-Y., Carden, M.J. and Trojanowski, J.Q. (1986) *J. Neurosci.*, **6**, 850–858.
- Lee, V.M.-Y., Carden, M.J., Schlaepfer, W.W. and Trojanowski, J.Q. (1987) *J. Neurosci.*, **7**, 3474–3488.
- Lee, V.M.-Y., Otvos, L., Carden, M.J., Hollosi, M., Dietzschold, B. and Lazzarini, R.A. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 1998–2002.
- Lehnert, M.E., Jorcano, J.L., Zentgraf, H., Blessing, M., Franz, J.K. and Franke, W.W. (1984) *EMBO J.*, **3**, 3279–3287.
- Leterrier, J.-F. and Eyer, J. (1987) *Biochem. J.*, **245**, 93–101.
- Levy, E., Liem, R.K.H., D'Eustachio, P. and Cowan, N.J. (1987) *Eur. J. Biochem.*, **166**, 71–77.
- 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.
- Maniatis, T., Fritsch, E.F. and Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
- Marchuk, D., McCrohon, S. and Fuchs, E. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 1609–1613.
- Miyatani, S., Winkles, J.A., Sargent, T.D. and Dawid, I. (1986) *J. Cell Biol.*, **103**, 1957–1965.
- Myers, M., Lazzarini, R.A., Lee, V.M.-Y., Schlaepfer, W.W. and Nelson, D.L. (1987) *EMBO J.*, **6**, 1617–1626.
- Napolitano, E.W., Chin, S.S., Colman, D.R. and Liem, R.K.H. (1987) *J. Neurosci.*, **7**, 2590–2599.
- Perler, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodner, R. and Dodgson, J. (1980) *Cell*, **20**, 555–566.

- Quax,W., van den Broek,L., Egberts,W.V., Ramaekers,F. and Bloemendal,H. (1985) *Cell*, **43**, 327–338.
- Rackwitz,H.-R., Zehetner,G., Frischauf,A.-M. and Lehrach,H. (1984) *Gene*, **30**, 195–200.
- 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.
- Slightom,J.L., Blechl,A. and Smithies,O. (1980) *Cell*, **21**, 627–638.
- Smithies,O. and Powers,P.A. (1986) *Phil. Trans. R. Soc., London Ser. B.*, **312**, 291–302.
- 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.
- Soares,M.B., Turken,A., Ishii,D., Mills,L., Episkapou,V., Cotter,S., Zeitlin,S. and Efstratiadis,A. (1986) *J. Mol. Biol.*, **192**, 737–752.
- Steinert,P.M. and Parry,D.A.D. (1985) *Annu. Rev. Cell Biol.*, **1**, 41–65.
- Steinert,P.M. and Roop,D.R. (1988) *Annu. Rev. Biochem.*, **57**, 593–625.
- Sternberger,L.A. and Sternberger,N.H. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 6126–6130.
- Wiborg,O., Hyldig-Nielsen,J.J., Jensen,E.O., Paludan,K., Marcher,K.A. (1982) *Nucleic Acids Res.*, **10**, 3487–3494.
- Yashihara,C.M., Lee,J.-D., and Dodgson,J.B. (1987) *Nucleic Acids Res.*, **15**, 753–770.
- Zakut,R., Shani,M., Givol,D., Neuman,S., Yaffe,D. and Nudel,V. (1982) *Nature*, **298**, 857–858.
- Zopf,D., Hermans-Borgmeyer,I., Gundelfinger,E.D. and Betz,H. (1987) *Genes Dev.*, **1**, 699–708.

Received on March 14, 1988; revised on April 25, 1988