The complete amino acid sequence of the major mammalian neurofilament protein (NF-L)

Norbert Geisler, Uwe Plessmann and Klaus Weber*

Max Planck Institute for Biophysical Chemistry, D-3400 Goettingen, FRG

Received 28 January 1985

The first complete amino acid sequence of a neurofilament protein has been established. Porcine NF-L contains 548 residues corresponding to a molecular mass of ~62 kDa. This value is noticeably smaller than the 68–72 kDa estimates from gel electrophoresis. Sequence comparison among the 6 non-epithelial intermediate filament (IF) proteins of warm-blooded vertebrates shows that the three NF proteins are the most remote members. Additionally and unexpectedly they reveal among each other lower sequence identity than the three non-neuronal IF proteins GFAP, desmin, and vimentin where the last two are particularly closely related. Certain schemes of IF protein evolution are discussed.

Neurofilament *Intermediate filament* *Keratin* *Amino acid sequence* *Evolution* *Desmin*

1. INTRODUCTION

Mammalian neurofilaments (NF) usually contain three proteins. L, M and H have apparent molecular masses of 68, 160 and 200 kDa in gel electrophoresis. Biochemical and partial sequence data show that in spite of the high molecular masses all three proteins [1–3] follow the common structural organization of intermediate filament (IF) proteins: a structurally conserved α-helical region of about 310 residues (rod) flanked by highly variable non-α-helical domains, the amino-terminal head- and the carboxy-terminal tailpiece [4–7]. The extra mass of NF proteins arises from long tailpiece extensions of high charge density, unusually rich in glutamic acid and lysine. This feature distinguishes NF proteins from all other IF proteins [1]. The first complete amino acid sequence of an NF protein reported here shows unexpectedly that the three NF proteins are less closely related among each other than the three non-neuronal IF proteins desmin, vimentin, and GFAP.

2. MATERIALS AND METHODS

L-protein from porcine spinal cord was treated with 2-nitro-5-thiocyanobenzoic acid [1,8]. DEAE chromatography in urea provided the amino-terminal 35-kDa fragment as the most basic part. The mixture of CNBr fragments of this domain was passed in urea through CM cellulose to provide fragment 1. The flow through fraction was gel filtered through Sephadex G-100 in NH_4HCO_3. The first main peak was subjected to preparative gel electrophoresis to separate fragment 2 from an overlap. Fragments 4 and 6 present in the second peak were separated in urea by DEAE chromatography, as 4 passes through whereas 6 is bound. The final peak of the G-100 column contained fragments 3 and 5, which were separated by HPLC. Enzymatic digests, preparative fingerprints, peptide recovery, amino acid analysis and stepwise Edman degradation were as in [1]. Residues 127–137, 221–241, and 273–283 were additionally verified by a gas phase sequenator. We thank Drs J. Vandekerckhove and J. Van Damme for this information.

* To whom correspondence should be addressed

Published by Elsevier Science Publishers B.V. (Biomedical Division)
00145793/85/$3.30

3. RESULTS AND DISCUSSION

Previous data provided the first 82 and the last 270 residues of porcine L [1,8]. To close the gap of about 200 residues we used the single cysteine residue. Treatment with 2-nitro-5-thiocyanobenzoic acid yielded two fragments in addition to uncleaved material [8]. The N-terminal 35-kDa fragment was easily separated from the other species on DEAE-cellulose as it lacks the very acidic tailpiece. The six CNBr fragments generated from 35 kDa were purified (see section 2). Fragment 1 (residues 1–63) was known from the sequence of the headpiece domain (residues 1–81) isolated as a proteolytic fragment [1]. Fragment 2 (15 kDa) extended the sequence from residue 64 to 191. The sequence was deduced from proteolytic peptides. These were separated by fingerprints, recovered and characterized by composition and stepwise degradation. Fragment 6 was found to start 5 residues prior to the sole tryptophan previously given as first residue of the C-terminal 270 residues [1,8]. The remaining gap was closed by fragments 3–5 which were sequenced as above. The relative order of fragments 3–5 relies on the common structural principles of all IF proteins (see above) and the excellent alignment with M in this region (fig.1).

The sequence proposal shows a polypeptide of 548 residues with a molecular mass close to 61.9 kDa. The higher values of 68–72 kDa seen by gel electrophoresis arise from the unusually highly charged tailpiece and this divergence seems even more pronounced for M and H [9]. No evidence for sequence heterogeneity was observed indicating a single L-protein species. Hybridization data using cloned cDNA put the gene number at 1 as well [10]. Here, the C-terminal 302 residues are given for murine L. In line with previous proposals [5] a number of exchanges in the tailpiece region are seen, whereas the murine and porcine proteins differ only by 2 conservative exchanges in the 160 residues of the rod currently established for murine L.

Of the 6 non-epithelial IF proteins, 3 are now known by full sequence: desmin (chicken) [5], vimentin (hamster) [11], and NF-L (pig). DNA and protein data on murine and porcine GFAP together document a fourth prototype [12,13]. For porcine M the first 436 residues are known and fragments are available for H [2,3]. Thus the structurally conserved rod domains (fig.1) of 5 of the 6 non-epithelial IF proteins can be compared over the full length for sequence identity (table 1). As predicted [11,13,14], desmin and vimentin are, with an identity of 73%, the closest pair, although an avian and a mammalian protein are compared. GFAP shares only 63% identity with desmin or vimentin. Interestingly, NF-L and NF-M are not only the most remote members (~50% identity in comparison with desmin and vimentin) but unexpectedly they have very strongly diverged from each other as they share only 54% identity. Essentially similar results are obtained when comparison is restricted to the rod segment established for H. Absolute values are, however, lower since the consensus-type sequences at the ends of the rod [1,5] are not covered. Again the three non-neuronal proteins are very much closer (60–68% identity) than the three NF proteins, which have noticeably diverged (L/M 46%; M/H 44%; L/H 37% identity). The relative relations seen for the rod domains seem paralleled by the divergence of the terminal domains. The three non-neuronal proteins have closely related tailpieces but GFAP shows a quite different headpiece [12,13]. The unusually charged tailpieces and their increasing length is found only in NF proteins [1–3].

Partial data on the rod exist also for porcine and hamster desmin [8,14,15], murine L [10], and porcine GFAP [13]. A few conservative exchanges between the rodent and porcine proteins are noticed. The carboxyterminal 122 residues of the desmin rod are known for 3 species. Seven of the 8 replacements seen earlier between porcine and chicken desmin [8,14] are also present in hamster desmin [15]. Thus the two mammalian proteins are much closer to each other (99.2%) than to the avian protein (93.4%). This species-specific drift together with the presence of separate genes for desmin and vimentin in all classes of vertebrates [15] favour a model where the corresponding genes arose by gene duplication from a precursor prior to the evolution of vertebrates. In such a model GFAP would have separated earlier from desmin/vimentin and the NF proteins diverged even earlier from the putative tree of IF precursors. Firm biochemical evidence for IF proteins in invertebrates exists only for NF, which have been characterized in *Myxicola* (160 kDa and possibly

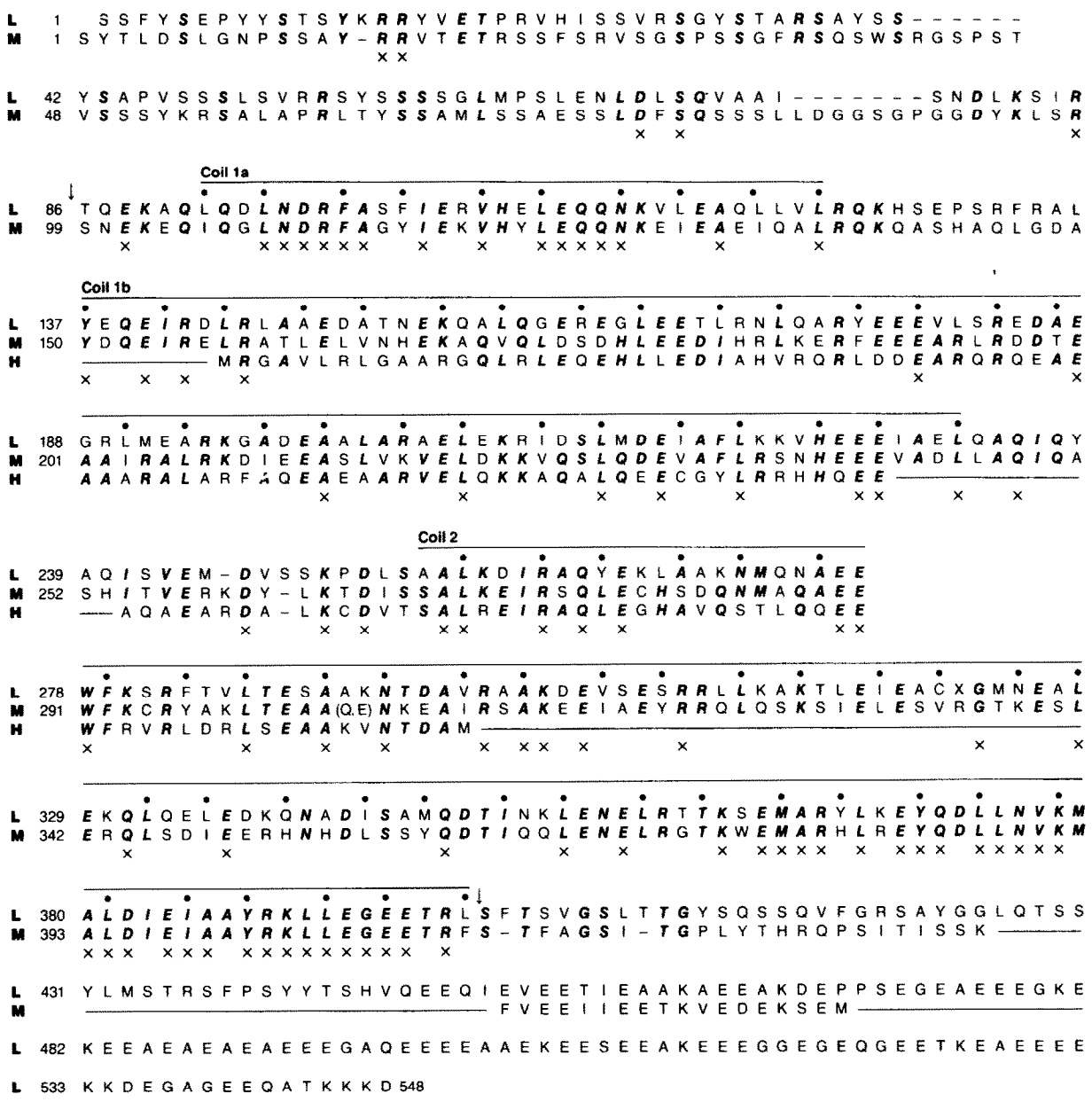

Fig.1. The sequence of L and comparison with the partial sequences established for the two higher molecular mass NF proteins M and H [2,3]. Identical residues present in at least two NF proteins are given by bold letters. Horizontal lines indicate not yet known sequences in M and H. Sequence arrangement is as discussed [5]. Coiled-coil forming α-helices are marked by lines and points indicate the primarily hydrophobic a and d positions of consecutive heptades. Crosses indicate residues identical in all non-epithelial IF proteins (for references to sequences of desmin, vimentin and GFAP see text). Note the consensus type sequences early in coil 1a and at the end of coil 2 [1,5]. The single X in the sequence of L is either lysine or arginine [1]. Due to a typographical error residue 6 of L was erroneously given as Q rather than E [1]. Dashes allow a better alignment in the non-α-helical regions. Along the rod (marked by arrows) length variability is very small. Between coil 1a and 1b (spacer 1) there are 13 residues in L and M and 11 in the 3 non-neuronal proteins. These show 21 residues between coil 1b and 2 (spacer 2) whereas L and M have 22. A slightly different presentation of IF proteins [6] proposes a further non-α-helical spacer some 15 residues past the central tryptophan (W). No length variability is seen in this region in any of the currently known IF proteins.

Table 1

Sequence identity (in %) along the entire rod of 5 distinct non-epithelial IF proteins

	V	D	G	L	M
V	0	73	63	53	50
D	(68)	0	63	53	49
G	(61)	(60)	0	48	48
L	(51)	(54)	(45)	0	54
M	(39)	(39)	(37)	(46)	0
H	(37)	(37)	(37)	(37)	(44)

V, D, and G are vimentin, desmin, and GFAP. The NF proteins are L, M, and H. Values in parentheses arise when comparison is made for all 6 proteins using only the segment established for H. See text and fig.1

150 kDa), *Loligo* (60 and 200 kDa), and *Aplysia* (60 and 65 kDa) [16–18]. Although these results indicate already multiple NF proteins and the presence of high molecular mass NF proteins, the general situation is more complicated because of a virtual lack of studies on other IF proteins in invertebrates (for a possible non-neuronal IF protein in *Drosophila* see [19]) and the number of NF proteins seen in vertebrates [20]. L, M and H are found only in mammals and birds. Reptiles and lower vertebrates while displaying an L-protein immunologically related to the mammalian L-protein, have only one high molecular mass protein. Thus, if a precursor gene of H and M arose during reptile evolution, additional influences must account for their later strong divergence, since desmin and vimentin are still closely related even though they may have separated already at or prior to vertebrate evolution. Further molecular information may answer some of the questions concerning IF proteins during evolution and may also connect the subclass of non-epithelial proteins with the two subclasses of epithelial keratins defined by sequences [6]. These are so far only known for vertebrates. Here they seem coded by at least 20 genes if keratin data on mammals are generalized [6,21].

ACKNOWLEDGEMENTS

We thank the Deutsche Forschungsgemeinschaft for financial support and A. Gruber and S. Gaspar for expert technical assistance.

REFERENCES

[1] Geisler, N., Kaufmann, E., Fischer, S., Plessmann, U. and Weber, K. (1983) EMBO J. 2, 1295–1302.
[2] Geisler, N., Fischer, S., Vandekerckhove, J., Plessmann, U. and Weber, K. (1984) EMBO J. 3, 2701–2706.
[3] Geisler, N., Fischer, S., Vandekerckhove, J., Van Damme, J., Plessmann, U. and Weber, K. (1985) EMBO J. 4, 57–63.
[4] Geisler, N., Kaufmann, E. and Weber, K. (1982) Cell 30, 277–286.
[5] Geisler, N. and Weber, K. (1982) EMBO J. 1, 1649–1656.
[6] Hanukoglu, I. and Fuchs, E. (1983) Cell 33, 915–924.
[7] Steinert, P.M., Rice, R.H., Roop, D.R., Trus, B.L. and Steven, A.D. (1983) Nature 302, 794–800.
[8] Geisler, N., Plessmann, U. and Weber, K. (1982) Nature 296, 448–450.
[9] Kaufmann, E., Geisler, N. and Weber, K. (1984) FEBS Lett. 170, 81–84.
[10] Lewis, S.A. and Cowan, N.J. (1985) J. Cell Biol., in press.
[11] Quax, W.J., Egbert, W.V., Hendriks, W., Quax-Jeuken, Y. and Bloemendal, H. (1983) Cell 35, 215–223.
[12] Lewis, S.A., Balcarek, J.M., Krek, V., Shelanski, M. and Cowan, N.J. (1984) Proc. Natl. Acad. Sci. USA 81, 2743–2746.
[13] Geisler, N. and Weber, K. (1983) EMBO J. 2, 2059–2063.
[14] Geisler, N. and Weber, K. (1981) Proc. Natl. Acad. Sci. USA 79, 4120–4123.
[15] Quax, W., Van Den Heuvel, R., Egberts, W.V., Quax-Jeuken, Y. and Bloemendal, H. (1984) Proc. Natl. Acad. Sci. USA 81, 5970–5974.
[16] Eagles, P.A.M., Gilbert, D.S. and Maggs, A. (1981) Biochem. J. 199, 89–100.
[17] Lasek, R.J., Krishnan, N. and Kaiserman-Abramof (1979) J. Cell Biol. 82, 336–346.
[18] Lasek, R.J., Oblinger, M.M. and Drake, P.F. (1983) Cold Spring Harbor Symp. Quant. Biol. 48, 731–744.
[19] Walter, M.F. and Biessmann, H. (1984) J. Cell Biol. 99, 1468–1477.
[20] Shaw, G., Debus, E. and Weber, K. (1984) Eur. J. Cell Biol. 34, 130–136.
[21] Moll, R., Franke, W.W., Schiller, D.L., Geiger, B. and Krepler, R. (1982) Cell 31, 11–24.