The structure of the human gene encoding protein gene product 9.5 (PGP9.5), a neuron-specific ubiquitin C-terminal hydrolase

Ian N. M. DAY, Lesley J. HINKS and R. J. THOMPSON

University Department of Clinical Biochemistry, Level D, South Laboratory and Pathology Block, Southampton General Hospital, Tremona Road, Southampton SO9 4XY, U.K.

Database search using a bovine thymus ubiquitin C-terminal hydrolase sequence indicated 54 % sequence identity with the abundant human neuron-specific protein gene product 9.5 (PGP9.5), which was then shown to possess the same activity [Wilkinson, Lee, Deshpande, Duerksen-Hughes, Boss & Pohl (1989) *Science* **246**, 670–673]. A yeast counterpart of the enzyme is also known. The human PGP9.5 gene, described here, spans 10 kb, contains nine exons and displays 5' features some common to many genes and some common with neurofilament neuron-specific enolase and Thy-1-antigen gene 5' regions.

INTRODUCTION

Ubiquitin is implicated in the regulation of a wide variety of cellular functions, but its most general role may be to tag damaged cytosolic proteins for degradation [1]. Covalent attachments between ubiquitin and other proteins or peptides involve the C-terminus of ubiquitin and α -amino and lysine ϵ -amino groups of the conjugate. A new enzyme family has recently been defined: the 'ubiquitin C-terminal hydrolases' (UCHs) [2]. By definition these enzymes hydrolyse small C-terminal adducts of ubiquitin to (re)generate the ubiquitin monomer. *In vitro*, various substrates have been used, although the natural substrate(s) are not yet known. The enzymes appear to be thiol proteinases. Several proteins that remain independent during column fractionations display activity, and those from bovine calf thymus have been named L1, L2, L3 and H2 [2]. L3 is the major activity in thymus, and its cDNA has been isolated. Database search revealed 54 % sequence identity with the neuron-specific protein gene product 9.5 (PGP9.5) [3], which was subsequently shown to be identical with L1, a minor activity (< 1 %) in thymus. A comparable yeast enzyme [4] also shows identity with these sequences.

UCH-L3 is abundant in calf thymus and resembles a form found in rabbit reticulocytes and rat liver [2,5]. The cellular specificity of PGP9.5 (UCH-L1) has been studied in detail by immunochemical methods [6–9]. It is highly specific to neurons and to cells of the diffuse neuroendocrine system and their tumours. It is present in almost all neurons, and averages 1–5 % of total soluble brain protein [10]. It behaves as a monomer of approx. 24 kDa and does not appear to have any associated factors [6,11; I. N. M. Day & R. J. Thompson, unpublished work]. The human cDNA has been described previously [12], and it appears that PGP9.5 expression is regulated at the transcriptional level. Disturbances of ubiquitin turnover [13] and the ubiquitination of neurofibrillary tangles found in Alzheimer's dementia and other neurodegenerative disorders [14] render the PGP9.5 gene and its product of considerable clinical interest. Additionally, since PGP9.5 is one of the most abundant products specific to, and present in, almost all neurons, its gene may exhibit powerful neuron-specific enhancer elements. Here we describe the analysis of the human PGP9.5 gene.

MATERIALS AND METHODS

Gene isolation, sequencing and analysis

A human genomic library in the vector λ 2001 was screened using human PGP9.5 cDNA labelled by random priming as described previously [15]. One clone was analysed in detail by restriction mapping and Southern blotting, using cDNA subfragments and subcloning into plasmid pUC18 by standard methods [16]. All DNA sequencing was performed by using the Sanger method [17].

Sequencing templates were either single-stranded DNA rescued from the phagemid pBS [18] or double-stranded pBS denatured by alkali [19]. Restriction-enzyme-directed and sonication shotgun [20] clones were created. Standard primers were used [23] in addition to gene-specific primers (Fig. 1b). The latter were synthesized by the β -cyanoethyl phosphoramidite method using an Applied Biosystems 381A synthesizer. Computer analysis was performed using the Staden software package supplied by Amersham International, implemented on a suitable personal computer.

RNA isolation and 5' primer extension

Samples of cerebral cortex from human post-mortem (24–36 h) brain were frozen at -70°C until use. Total RNA was extracted by ultracentrifugation through a CsCl cushion after tissue homogenization in guanidinium thiocyanate [21]. PGP9.5 mRNA was affinity-selected on PGP9.5 cDNA (subfragment B; see Fig. 1a) immobilized on nylon membranes, and RNA integrity was checked by Northern blotting (using subfragment B as a probe) after formaldehyde/gel electrophoresis [21]. 5' Primer extension was performed using an end-labelled 17-mer oligonucleotide CAGCACTTTGTTTCAGCA complementary to a region near the 5' end of the mRNA (Fig. 2a). Oligonucleotide annealing and murine-leukaemia-virus reverse-transcriptase reaction were performed on both total and affinity-selected RNA under conditions as described [21], in the presence of 10 mM-vanadyl ribonucleoside complexes. The labelled extension products were sized against a sequencing ladder.

Abbreviations used: PGP9.5, protein gene product 9.5; UCH, ubiquitin C-terminal hydrolase.

These sequence data will appear in the EMBL/GenBank/DBJ Nucleotide Sequence Databases under the accession number X17377.

Fig. 1. (a) Scale map of the human PGP9.5 gene and (b) relationship of the exons to the previously published protein sequence

(a) Exons are displayed as blocks on the top line with the 5' end of the gene on the left. A_i, B_i and C_i designate the three regions (divided by two *Eco*RI sites) of a λ 2001 genomic clone which were subcloned into pUC18 for further analysis. Since the only two *Eco*RI sites present in the gene are in exons, A_i, B_i and C_i represent the hybridization targets for three cDNA regions designated A, B and C similarly divisible by the same two *Eco*RI sites (see also the Results section). Useful restriction sites in A_i and B_i are indicated as follows: H, *Hind*III; K, *Kpn*I; P, *Pst*I; RI, *Eco*RI; S, *Sma*I; and X, *Xba*I. Gene-specific 17-mers representing the 5' end of the seventh exon and the 3' end of the eighth exon were used initially to locate their approximate position by hybridization and then to define their respective 3' and 5' boundaries by sequencing. The *Xba*I fragment containing these exons was subjected to saturation shotgun cloning. (b) The protein sequence is displayed exon-by-exon in the one-letter amino acid code. Codons divided by a splice are indicated in lower-case letters, with numerical subscripts to indicate the disposition of the codon positions between exons. The apparent active site is indicated by a star, and positions identical with the thymic homologue UCH-3 are underlined.

RESULTS

One genomic clone containing targets for PGP9.5 cDNA subfragments A and C representing the 5' and 3' ends of the mRNA was analysed in detail. It contained only two restriction-*Eco*RI sites, both of which are present in the cDNA and which defined the cDNA subfragment nomenclature A, B and C (Fig. 1a). The genomic targets for A, B and C, called A_i, B_i and C_i, were subcloned into pUC18, and minimal restriction maps and exon-hybridization maps were constructed in pUC18. Exon boundaries were determined by sequencing, and in every instance obligatory AG and GT splice sites [22] with flanking consensus were found. The map is shown in Fig. 1(a), and the relationship of the splice sites to the enzyme sequence is shown in Fig. 1(b). The nine exons span 10 kb of genomic DNA. cDNA probe B has been used previously to show that the gene is single copy [15], and various blots indicate correspondence of

restriction fragments in this genomic clone with those identified in total genomic DNA by Southern blotting (results not shown).

The 5' end of the cloned insert was sequenced to completion. Relative to the cDNA a short nucleotide inversion of sequence was identified, indicating a 5' artefact in the original cDNA clones (see the Discussion section). Complementary 5' primer extension using the mRNA as template produces a single major product of 101 nucleotides (Fig. 2b). Relative to the genomic sequence, this places the transcriptional initiation site at a G nucleotide 27 nucleotides 3' to the only TATA sequence in the region. The 5' end is not depleted of the dinucleotide CpG, and displays several direct repeats (Fig. 2a). Direct comparison of this region with known regulatory elements, and with the equivalent regions in neurofilament [23–26], neuron-specific enolase [27] and Thy-1 [28] genes indicates a number of common features and elements discussed below.

Fig. 2. (a) Sequence of the 5' end of the human PGP9.5 gene and (b) mapping the 5' end of PGP9.5 mRNA

(a) The TATA region is enclosed in a box, and the transcriptional start site (see b) is shown by a downward-pointing arrow. The region indicated by the wavy line has been observed in inverted form in cDNA clones ([12]; I. N. M. Day & R. J. Thompson, unpublished work). Splice junctions are displayed by a solidus and adjacent underlining, and the coding sequence is accompanied by one-letter amino acid code above the line. A *Pst*I site useful in achieving ordered subcloning from this region is shown. Single-stranded templates spanning this region 5' → 3' and 3' → 5', and *Pst*I → 5' and *Pst*I → 3' clones enabled sequencing of both strands, with overlaps. Four direct repeats upstream of the TATA box are underlined in numbered pairs, and other features of this region of sequence are discussed in the text. The 17-mer used for mapping the 5' end of the mRNA (see b) has an A as its most 3' nucleotide; this A is complementary with the T of the methionine codon found in the second exon (i.e. MLNK). (b) The ³²P-end-labelled oligonucleotide CAGCACTTTGTTTCAGCA (relationship to coding sequence described in the legend to a above) was used to prime first-strand synthesis using PGP9.5 mRNA as described in the text. Even using total brain RNA, only a single major extension product of 101 nucleotides is obtained, placing the 5' end of the mRNA at the position marked by the arrow in a. (i) Nucleotide sequencing ladder; (ii) PGP9.5 mRNA 5' end analysis.

DISCUSSION

At the protein level, the presence of three closely spaced potential initiator methionine residues was not recognized from the cDNA, owing to a short 5' inversion attributable to the cloning method [12,21]. The size and amino acid content of PGP9.5 [8] do not adequately identify the translational start site, and the *N*-terminus is blocked from sequencing [8]. However, cryptic 5' ATG sites are rare in mRNAs [29], and a trace CNBr fragment of initial manual sequence Glu-Leu/Ile has been identified [11] and the comparable UCH-3 and yeast enzyme display conservation of the first of these methionine residues [3]. Therefore it is likely that the translated protein is as shown in Fig. 1(b). Exon 2 is only 12 nucleotides in length, but this and the other exon/intron lengths, and number, fall within known statistical limits [30]. It is notable that the longest coding exon contains both the apparent active-site thiol and the longest conserved domain [2]. It will be of interest to determine the

interactions of this domain with ubiquitin, and its conjugates, and to determine whether this domain carries any more fundamental evolutionary or biological significance in the multiple and extensively diversified pathways regulating and regulated by ubiquitin [1].

At the DNA level, the human PGP9.5 gene displays features typical of mammalian genes, including obligatory intron splice sequences [22], transcriptional initiation at a purine residue and a TATA box [34] at position -20 to -30 relative to the transcription start site, although there is no CCAAT box [31] near position -70.

At the technical level we note that undegraded PGP9.5 mRNA (judged by Northern blotting) is obtained from 24 h post-mortem brain. *In vitro* translation and two-dimensional protein electrophoresis have shown that RNA of adequate quality is obtained from similar material [32]. Therefore we have no reason to doubt the validity of the primer-extension data: additionally, these data cannot be confounded by the presence of a 5' non-

coding exon [33], since there is not suitably placed splice acceptor site downstream of the apparent transcriptional start site but upstream of sequence already known to be present in cDNA. The region 5' to the TATA box in the PGP9.5 gene is not depleted of the dinucleotide CpG, which is atypical in the human genome [33]. This may reflect an unmethylated 5' island typical of vertebrate genes [33]. It is thought that methylation at CpG, which renders it hypermutable, is hindered by regulatory and transcriptional factors binding the 5' ends of genes [33]. Screening of this region of the PGP9.5 gene for sequence elements representing binding sites for such factors indicates several direct repeats typical of such regions [27], which may reflect multiple binding sites or specialized DNA structures. One repeat (numbered 2 in Fig. 2a) is similar to the β -globin CACCC box [31]. Comparable elements are present in the rat neuron-specific enolase gene [27], human neurofilament heavy subunit [14] and mouse neurofilament medium subunit [17] genes. 5' Regions proximal to the TATA box of at least 200 nucleotides have been published for these genes and also for the genes encoding human neurofilament medium subunit [23], human neurofilament light subunit (only 100 nucleotides) [25] and mouse Thy-1 [28] genes. Except for Thy-1, which has alternative initiation sites 260 bp apart and is also expressed in lymphocytes, each one displays a true TATA box with several flanking A and T nucleotides, in contrast with a number of 'housekeeping' genes (see [28]). Whether there is tissue-specific variation of the pre-initiation complex which includes the TATA-binding transcription factor TFIID [34] is yet unknown. Neither the PGP9.5 gene nor these other neuronally expressed genes display the CCAAT box [31] commonly found at position -75, although each displays a CC₆TT₆C motif at approximately this site. The Thy-1 gene lacks TATA and CCAAT boxes, but the alternative initiation sites both display CTCCTGCT at position -48. The PGP9.5 gene contains two similar elements, CCCCCCTGCA at -66 and GCCCCTGCT, at -75, as does the rat neuron-specific enolase gene (CCGCCTGCA) at a comparable position (-64, -78 or -85 relative to its multiple transcriptional start sites). The sites in the Thy-1 gene are known to be essential and sufficient for its transcription in lymphoid cell lines, and they probably operate in neurons as well [28]. It is conceivable, therefore, that these elements are the targets for members of the differentially regulated families of CCAAT-binding proteins already known [34]. Direct analysis of the regulation of the proximal 5' region by cellular transfection assays, genomic walking to isolate distal elements, and investigation of the role of this gene in the events of neuronal ubiquitination and neurofibrillary tangle formation are made possible by the work presented here.

Dr. T. Rabbitts is thanked for providing the genomic library, Mrs. P. Allsopp and Mr. G. Vautier for technical assistance at the outset of the project, Dr. Paul A. Whittaker for helpful discussion, Dr. Mark Pickett for oligonucleotide synthesis, and Dr. Keith Wilkinson for communicating results before publication. This work was largely supported by a grant from the Wellcome Trust to R.J.T. Work in the

laboratory of I.N.M.D. is supported by the Medical Research Council of Great Britain, The British Epilepsy Research Foundation and the Sir Halley Stewart Trust. Mrs. Wendy Pringle is thanked for typing the manuscript.

REFERENCES

1. Hershko, A. (1988) *J. Biol. Chem.* **263**, 15237-15240
2. Mayer, A. N. & Wilkinson, K. D. (1989) *Biochemistry* **28**, 166-172
3. Wilkinson, K. D., Lee, K., Deshpande, S., Duerksen-Hughes, P., Boss, J. M. & Pohl, J. (1989) *Science* **246**, 670-673
4. Miller, H. I., Henzel, W. J., Ridgway, J. B., Kuang, W. J., Chisholm, V. & Liu, C. C. (1989) *Biotechnology* **7**, 698-704
5. Pickart, C. M. & Rose, I. A. (1985) *J. Biol. Chem.* **260**, 7903-7910
6. Doran, J. F., Jackson, P., Kynoch, P. A. M. & Thompson, R. J. (1983) *J. Neurochem.* **40**, 1542-1547
7. Thompson, R. J., Doran, J. F., Jackson, P., Dhillon, A. P. & Rode, J. (1983) *Brain Res.* **278**, 224-228
8. Thompson, R. J. & Day, I. N. M. (1988) in *Neuronal and Glial Proteins* (Marangos, P. J., Campbell, I. C. & Cohen, R. M., eds.), vol. 2, pp. 209-228, Academic Press, London and New York
9. Rode, J., Dhillon, A. P., Doran, J. F., Jackson, P. J. & Thompson, R. J. (1985) *Histopathology* **9**, 147-158
10. Jackson, P. & Thompson, R. J. (1981) *J. Neurol. Sci.* **49**, 429-438
11. Day, I. N. M. (1987) Ph.D. Thesis, University of Cambridge
12. Day, I. N. M. & Thompson, R. J. (1987) *FEBS Lett.* **210**, 157-160
13. Gallo, J. M. & Anderton, B. H. (1989) *Nature (London)* **337**, 687-688
14. Mori, H., Kondo, J. & Ihara, Y. (1987) *Science* **235**, 1641-1644
15. Day, I. N. M., Allsopp, M. T. E. P., Moore, D. C. McN. & Thompson, R. J. (1988) *Biochem. Soc. Trans.* **16**, 323
16. Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
17. M13 Handbook, Amersham International, Amersham
18. Short, J. M., Fernandez, J. M., Sorge, J. A. & Huse, W. D. (1988) *Nucleic Acids Res.* **16**, 7583-7600
19. Hattori, M. & Sakaki, Y. (1986) *Anal. Biochem.* **152**, 232-238
20. Anderson, S. (1981) *Nucleic Acids Res.* **9**, 3015-3027
21. Berger, S. L. & Kimmel, A. R. (1987) *Guide to Molecular Cloning Techniques (Methods Enzymol. 152)*, Academic Press, London and New York
22. Shapiro, M. B. & Senapathy, P. (1987) *Nucleic Acids Res.* **15**, 7155-7174
23. Lees, J. F., Shneidman, P. S., Skuntz, S. F., Carden, M. J. & Lazzarini, R. A. (1988) *EMBO J.* **7**, 1947-1955
24. Myers, M. W., Lazzarini, R. A., Lee, V. Y., Schlaepfer, W. W. & Nelson, D. L. (1987) *EMBO J.* **6**, 1617-1626
25. Julien, J. P., Grosveld, F., Yazdanbakhsh, K., Flavell, D., Meijer, D. & Mushynski, W. (1987) *Biochim. Biophys. Acta* **909**, 10-20
26. Levy, E., Liem, R. K. H., D'Eustachio, P. & Cowan, N. J. (1987) *Eur. J. Biochem.* **166**, 71-77
27. Sakimura, K., Kushiya, E., Takahashi, Y. & Susuki, Y. (1987) *Gene* **60**, 103-113
28. Ingraham, H. A. & Evans, G. A. (1986) *Mol. Cell. Biol.* **6**, 2923-2931
29. Kozak, M. (1987) *Nucleic Acids Res.* **15**, 8125-8148
30. Hawkins, J. D. (1988) *Nucleic Acids Res.* **16**, 9893-9905
31. Myers, R. M., Tilly, K. & Maniatis, T. (1986) *Science* **232**, 613-618
32. Morrison, M. R. & Griffin, W. S. T. (1980) *Anal. Biochem.* **113**, 318-324
33. Bird, A. P. (1987) *Trends Genet.* **3**, 342-347
34. Mitchell, P. J. & Tjian, R. (1989) *Science* **245**, 371-377