Structure and Chromosome Assignment of the Murine p36 (Calpactin I Heavy Chain) Gene†

Patrick Amiguet,‡ Peter D'Eustachio,§ Torsten Kristensen,∥ Rick A. Wetsel,⊥ Chris J. M. Saris,#,○ Tony Hunter,# David D. Chaplin,△ and Brian F. Tack*,‡

Department of Immunology, Research Institute of Scripps Clinic, La Jolla, California 92037, Department of Biochemistry, New York Medical School, New York, New York 10016, Institute of Molecular Biology, University of Aarhus, Aarhus C, Denmark, Department of Pediatrics and of Microbiology and Immunology, Washington University School of Medicine, St. Louis, Missouri 63110, Molecular Biology and Virology Laboratory, The Salk Institute, San Diego, California 92318, and Department of Internal Medicine and Howard Hughes Medical Institute, Washington University School of Medicine, St. Louis, Missouri 63110

Received July 24, 1989; Revised Manuscript Received September 14, 1989

ABSTRACT: p36 is a major substrate of both viral and growth factor receptor associated protein kinases. This protein has recently been named calpactin I heavy chain since it is the large subunit of a Ca^{2+}-dependent phospholipid and actin binding heterotetramer. The primary structure of p36 has been determined from analysis of cloned cDNA. The protein contains 338 amino acids, has an approximate molecular weight of 39 000, and is comprised of several distinct domains, including four 75 amino acid repeats. From two overlapping cosmid clones isolated from different mouse genomic liver libraries, the complete intron/exon structure of the p36 gene was determined and the 5′ and 3′ noncoding regions of the gene were analyzed. The coding and 3′ untranslated region of the p36 gene contains 12 exons which range in size from 48 to 322 base pairs (bp) with an average size of 107 bp. The repeat structures found at the protein level are not delineated by single exons, but the N-terminal p11-binding domain is encoded by a single exon. Structural mapping of the gene demonstrated that the lengths of the first two introns in the coding region are together approximately 6 kilobases (kb), while the other introns range in size from 600 to 3600 bp with an average size of 1650 bp. The p36 gene is at least 22 kb in length and has a coding sequence of approximately 1 kb, representing only 4.5% of the gene. Analysis of mouse X Chinese hamster hybrid cell lines, together with typing of a restriction fragment length polymorphism in recombinant inbred strains of mice, localized the p36 gene (*Call h*) to chromosome 9, midway between *Apoa1* and *d*.

p36, also termed calpactin I heavy chain, is an approximately 39-kDa protein that has previously been shown to be a major substrate for protein-tyrosine kinases [Radke & Martin, 1979; Erikson & Erikson, 1980; Radke et al., 1980; Cheng & Chen, 1981; Cooper & Hunter, 1981a,b, 1983a; Decker, 1982; Nakamura & Weber, 1982; Greenberg & Edelman, 1983a,b; for reviews see Cooper and Hunter (1983b) and Hunter (1988)]. Subcellular fractionation and immunolocalization studies have demonstrated the association of p36 with the plasma membrane and the cortical skeleton underlying the membrane (Cooper & Hunter, 1983a; Courtneidge et al., 1983; Greenberg & Edelman, 1983a; Lehto et al., 1983; Nigg et al., 1983; Radke et al., 1983; Gould et al., 1984; Zokas & Glenney, 1987). In a variety of cell types p36 codistributes with nonerythroid spectrin, a cytoskeletal protein that is localized to the cortical submembraneous skeleton (Greenberg & Edelman, 1983a; Lehto et al., 1983; Gerke & Weber, 1984; Gould et al., 1984). However, p36 is absent from erythrocytes (Gould et al., 1984). In vitro studies have demonstrated specific Ca^{2+}-dependent interactions between intestinal p36 and both F-actin (Gerke & Weber, 1984; Glenney, 1986a) and nonerythroid spectrin (Gerke & Weber, 1984). Microinjection of antibodies directed against nonerythroid spectrin, however, has no effect on the distribution of p36 (Mangeat & Burridge, 1984), suggesting that nonerythroid spectrin is not tightly associated with p36 in vivo. Futhermore, since p36 does not colocalize with F-actin in cells (Nigg et al., 1983; Radke et al., 1983) and its distribution is not affected by drugs that disrupt actin filaments (Alitalo et al., 1984), p36 is apparently not tightly associated with F-actin in the cell either. Instead, the membrane localization of p36 may be explained by its ability to bind to anionic phospholipids in a Ca^{2+}-dependent manner (Glenney, 1985, 1986a,b; Drust & Creutz, 1988; Glenney & Zokas, 1988). This possibility is supported by the observation that the bulk of p36 can be released from cellular particulate fractions with Ca^{2+}-chelating agents (Cooper & Hunter, 1983a; Courtneidge et al., 1983; Greenberg & Edelman, 1983b; Gerke & Weber, 1984; Glenney & Tack, 1985). The p36 monomer has multiple Ca^{2+}/phospholipid binding sites and can bind up to four Ca^{2+} ions in the presence of phospholipid (Zokas & Glenney, 1988). Therefore, one would expect p36 to be able to promote membrane/membrane interaction, and this has been observed experimentally (Glenney et al., 1987; Drust & Creutz, 1988).

Calpactin I isolated from bovine intestine has two distinct subunits with molecular weights of 39 000 (p36, heavy chain) and 11 000 (p11,[1] light chain), respectively, which form a

† This is publication number 5422 IMM from the Research Institute of Scripps Clinic. This work was supported by Public Service Grants AI22214 and AI19222 (to B.F.T.), GM32105 (to P.D.), AI25022 (to R.A.W.), and CA39780 (to T.H.).

* To whom correspondence should be addressed.

‡ Research Institute of Scripps Clinic.

§ New York Medical School.

∥ University of Aarhus.

⊥ Department of Pediatrics and of Microbiology and Immunology, Washington University School of Medicine.

The Salk Institute.

○ Present address: Section of Molecular Genetics, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands.

△ Department of Internal Medicine and Howard Hughes Medical Institute, Washington University School of Medicine.

0006-2960/90/0429-1226$02.50/0

tetramer containing two heavy chains and two light chains (Gerke & Weber, 1984; Glenney & Tack 1985). Biochemical analysis of the p36 monomer shows that the four Ca^{2+}/phospholipid binding sites lie in a C-terminal protease-resistant core of about 33 000 (Glenney, 1986b), which is linked to a short N-terminal domain by a flexible region that is sensitive to a variety of proteases (Glenney & Tack, 1985; Johnsson et al., 1986). p11 binds to the N-terminal domain of p36 (Glenney & Tack, 1985; Johnsson et al., 1986). p11, which is related to the S100 family of Ca^{2+}-binding proteins (Glenney & Tack, 1985; Saris et al., 1987), exists as a dimer, and this linkage stabilizes the calpactin I tetramer (Glenney, 1986b). In the cell p11 is tightly linked to the cytoskeleton, and this interaction anchors the calpactin I tetramer to the cytoskeleton (Zokas & Glenney, 1988).

The primary structures of human, bovine, and murine p36 have been completely defined from sequence analysis of cDNA clones (Huang et al., 1986; Kristensen et al., 1986; Saris et al., 1986). The C-terminal domain of p36 contains four internal repeats of approximately 75 amino acids, leaving a distinct N-terminal domain of 30 amino acids. The sequence of p36 reveals that it belongs to a family of Ca^{2+}-binding proteins with related 75 amino acid repeats that all bind to cell membranes. This family includes the lipocortins, lipocortin I (= calpactin II = p35) and lipocortin II (= calpactin I = p36), and several other proteins [Geisow et al., 1986; Pepinsky et al., 1988; reviewed by Crompton et al. (1988) and Hunter (1988)]. The p36 N-terminal region of 12 amino acids interacts with p11 (Johnsson et al., 1988). Two phosphorylation sites are found at positions 23 (Tyr) and 25 (Ser), which are phosphorylated by a variety of protein-tyrosine kinases and protein kinase C, respectively (Glenney & Tack, 1985; Gould et al., 1986). The biochemically defined p36 core, which contains the four Ca^{2+}/phospholipid binding sites, clearly corresponds to the C-terminal region of the sequence starting at about residue 35 with its four tandem 75 amino acid repeats. This implies that each repeat functions as a Ca^{2+}/phospholipid binding site. The most highly conserved region of each repeat is a 17 amino acid consensus sequence first identified by Geisow et al. (1986), who noted this sequence in several proteins of this family. This 17 amino acid sequence is conserved in every repeat of all the proteins in this family and may represent a new type of Ca^{2+} and/or phospholipid binding site.

It has been postulated that exons represent genetic building blocks that code for structural or functional domains of proteins (Smith et al., 1979; Gilbert, 1978; Campbell & Porter, 1983; Craik et al., 1982). To determine whether the functional domains and internal repeats of p36 are organized in this fashion, we have examined the structural organization of the p36 gene and defined the exon arrangement. Our studies provide the framework for the further characterization of the diverse binding sites of the protein through deletion and expression experiments.

Experimental Procedures

Enzymes and Reagents. Restriction endonucleases were purchased from Promega Biotec and Boehringer Mannheim Biochemicals, a nick-translation kit was purchased from Bethesda Research Laboratories, and radiolabeled nucleotides were purchased from Amersham Corp.

Construction and Screening of the Cosmid Library. The cosmid libraries were constructed in the pTCF vector (Grosveld et al., 1982) as previously described (Chaplin et al., 1983) by using genomic DNA from either BALB/c or B10.D2/0Sn liver. Filters were prehybridized and hybridized at 42 °C in 50% formamide buffer. The two cosmid libraries were probed with two different ^{32}P-labeled p36 murine cDNA fragments representing positions 404–1137 and 64–907 of the full-length cDNA (Saris et al., 1986). Positive clones were colony-purified, and cosmid DNA was isolated (Maniatis et al., 1982).

Sequencing of Intron/Exon Junctions. Cosmid DNA was sequenced by using the "shotgun" method (Bankier & Barrell, 1983). Fragments of 300–600 bp in length, randomly generated by sonication, were ligated into the *Sma*I site of M13mp8. Exon-containing phage were identified by transferring a portion of the phage to GeneScreen filters (DuPont-New England Nuclear) in duplicate and probing the filter with the appropriate ^{32}P-labeled p36 cDNA fragments. These phage were then sequenced by the dideoxy chain termination method (Sanger et al., 1977) using α-^{35}S-labeled deoxyadenosine 5′-(thio)triphosphate (Biggin et al., 1983).

Southern Blot Analysis. Two micrograms of cosmid DNA were digested to completion with the appropriate restriction endonucleases and electrophoresed through a 1% agarose gel. The DNA was transferred to GeneScreen, and the filters were probed at 42 °C in 50% formamide buffer. M13 subclones, labeled by using a reaction mixture identical with that used in sequencing except that [α-^{32}P]dCTP was used as the labeled nucleotide and dideoxy nucleotides were omitted, were used as probes. After hybridizing overnight, the filters were washed in 0.2× SSC (1× = 0.15 M NaCl, 0.015 M sodium citrate, pH 7) and 0.1% sodium dodecyl sulfate at 65 °C and exposed to Kodak XAR film with an intensifying screen for 30 min–4 h at −70 °C.

Chromosomal Assignment. Mouse X Chinese hamster somatic hybrid cell lines were cultured, assayed for mouse chromosome content, and typed for the presence of mouse-specific genomic DNA sequences by Southern blotting as described previously (Saris et al., 1987). Mouse genomic DNA was isolated from liver nuclei and typed for restriction fragment length polymorphisms by Southern blotting as described previously (Saris et al., 1987) using a 3′ end fragment of p36 cDNA. All inbred and recombinant inbred mice used for genetic analysis were obtained from The Jackson Laboratory (Bar Harbor, ME).

Results

Murine p36 cDNA fragments were used to screen two mouse genomic cosmid libraries. Four clones were identified, and two of them, RW22 and B34, one from each library, were used to analyze the intron/exon junctions of the p36 gene and determine the intron lengths. Together, these clones span the entire p36 gene coding sequence. DNA from both clones was sonicated, and fragments between 300 and 600 bp were subcloned into M13mp8 bacteriophage. Clones hybridizing with ^{32}P-labeled p36 cDNA fragments were sequenced by the dideoxy chain termination method. DBAUTO and DBUTIL programs (Staden, 1980, 1982) on a VAX VMS computer were used to compile and align the sequences obtained. The consensus sequences were then compared to the cDNA sequence (Saris et al., 1986), and the intron/exon junctions were determined (Tables I and II). The p36 gene as defined by the cDNA was found to contain 12 exons. Beginning at the ATG start codon and ending at the polyadenylation signal, the exons average 107 bp and range from 48 to 322 bp in length. The first exon contains 48 bp of coding sequence including the ATG start codon (Figure 1A). The last exon (exon 12) contains 54 bp of coding sequence (Figure 1B), a TGA stop codon, and

[1] Abbreviations: p36, calpactin I heavy chain; *Call*h, p36 gene; p11, calpactin I light chain; p35, calpactin II; bp, base pair(s); kb, kilobase(s); cM, centimorgan.

A

TTTTTTTTTCAGCTTTTTTTTTCTTCAAAATGTCTACTGTCCACGAAATCCTGTGCAAGC
TCAGCCTGGAGGGTGATGTAAGT

B

CTGCAGGACACCAAGGGTGACTACCAGAAGGCACTGCTGTACCTGTGTGGTGGGGATGAC
TGAAGGGCTCAGCACAGTGGATCACCCAGAAGTGGCTCTACCTGTGCCCCAACCTGGCGT
TCTAGAGACTTCGCTCTCCACTAATGGACCCCTGAGCTCCTCCCTGTGAGGATGATGACA
GGGCTGCCGACCCTTTCCCCATCTTAGCTGCCCTTGCCTGGCTTTCTCCTCATTCTCTCC
TTTATGCCAAAGAAGTGAACATTCCAGGGAGTGGGGCGTCAGTCTGTGACATGAGACACT
TCCTCTTATGTACTGTGTCGTGAATAAACCGTTTTTACTTTAGAAACAAGAATGCCGTTG
CCTTTGCTTTCAAGCCACTGTTCAGAAAGCAGACATACTTGGGATTTGATTGTTCGCCCA
GGGAGTTAGGCCCCTAGAGGATGAAAGTTCTAGAACTTATCTCTTCCCTCTGGTAGCCAC
TATACAGATGGCACTTTTTTAAAAGGATGAGTGAATTTAAAATTATATACAGTTCAAATT
TTACCCCTTAGGCAAAGGGCTACAGCTCTTATC

FIGURE 1: Sequence of the 5′ (A) and 3′ (B) regions of the p36 gene: (heavy underline) coding sequence; (dashed underline) possible splice acceptor; (light underline) start or stop codon; (double underline) polyadenylation signal. 5′ and 3′ ends of clone 123 cDNA (Saris et al., 1986) are indicated by vertical arrows.

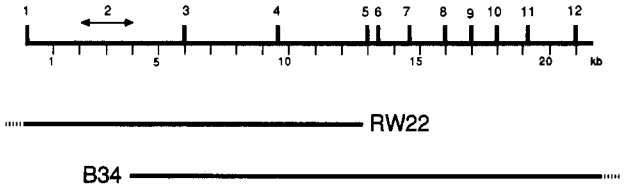

FIGURE 2: Map of the murine p36 gene. The positions of the 12 exons are indicated by the numbers above the line. The overlap of the two genomic clones relative to the p36 gene is indicated below. The position of exon 2 has not been identified precisely (probable range of its position is indicated by arrows).

a polyadenylation signal 259 bp downstream of the stop codon. No differences were found between the exonic sequences determined here for BALB/c genomic DNA and those obtained from the NIH3T3 cell p36 cDNA, even in the 3′ untranslated region (Saris et al., 1986).

The intron sizes were determined by Southern blot analysis. DNA from both cosmid clones was digested with various restriction enzymes. The bands were separated in 1% agarose gels, the DNA was transferred to nylon filters, and the blots were probed sequentially with the appropriate M13 subclones derived from the genomic clones (data not shown). Introns 1 and 2 were difficult to map with any confidence because of the multiple occurrence of many restriction endonuclease recognition sites. However, together they span a region of approximately 6 kb (Table I). The remaining introns average 1650 bp and range from 600 to 3600 bp in length. The p36 gene (Figure 2) covers a region of approximately 22 kb, with the first five exons spanning 13 kb and the last seven exons clustered in the last 8 kb of the gene. The coding sequence represents approximately 4.5% of the gene. Of 21 kb of intron,

Table I: p36 Intron and Exon Sizes

exon	length (bp)	intron	length (bp)
1[a]	48	I	6000[c]
2	100	II	
3	95	III	3600
4	114	IV	3300
5	91	V	600
6	80	VI	1200
7	60	VII	1350
8	94	VIII	1000
9	96	IX	900
10	59	X	1200
11	126	XI	1750
12[b]	322		

[a] From the ATG start codon. [b] Ending at the polyadenylation signal. [c] For introns I and II together.

Table II: Intron/Exon Junction Sequences

intron		exon					intron	
		ATG	1	GGT	GAT		gta	agt
ttg	cag	CAT	2	ACC	AAA	G	gta	ggc
cca	cag	GA	3	AAA	AAG		gta	ctg
tta	cag	GAG	4	ATG	AAG		gta	aac
tcc	cag	GGC	5	AAG	GAA	A	gtg	agt
cct	cag	TG	6	GCA	AAG		gtt	ggt
cta	cag	GGC	7	GCC	CGG		gta	tgt
cca	cag	GAG	8	CAG	AAA	G	gtg	ggt
tct	tag	TG	9	AAC	CTG	G	gtg	agc
cca	cag	TC	10	ATG	AAG		gta	agg
ttt	cag	GGC	11	CAG	CAA		gta	aag
ctg	cag	GAC	12	AAT	AAA			

approximately 6 kb were sequenced. The 5′ ends of the introns have the consensus sequence GTAAGT with a frequency of 100%, 100%, 64%, 55%, 73%, and 46% at each base, respectively. The sequence of the 3′ end of the introns was YYNYAG (Y represents a pyrimidine nucleoside, N represents any base) with a frequency of 100% at all positions (Table II). The 3′ end consensus sequence of the exon was AG, with a frequency of 82% and 73%, respectively. In 64% of the cases the exons started with a G.

p36 is composed predominantly of four repeats of 75 amino acids each. Figure 3 shows the intron/exon junction positions within these repeats. There is no striking correlation between the boundaries of the exons and the boundaries of the amino acid repeats. A DIAGON analysis of the coding sequence using a match of 44 of 99 bases was made to define the relationships between the exons (data not shown). Three of the four repeats found at the amino acid level can be seen at the nucleotide level. Exons 1, 2, and 12 are not found within the repeating units at the nucleotide level. Exons 3, 5, 8, and 11 share significant sequence homology (>50%), whereas the other exons show a more limited homology.

Southern blot analysis of *Eco*RI-digested mouse genomic DNA yielded a single fragment of 4.8 kb reactive with a 3′ end p36 cDNA probe at high stringency (0.2× SSC/65 °C), consistent with the existence of a single structural gene en-

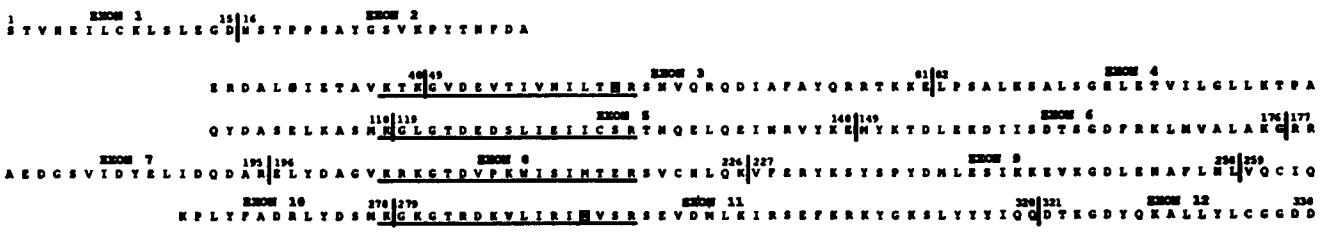

FIGURE 3: Murine p36 amino acid sequence. The four C-terminal 75 amino acid repeats are aligned, and the 17 amino acid consensus sequence identified by Geisow et al. (1986) is underlined. The vertical lines indicate the positions of the introns; amino acid residue numbers are given on either side of these exon/exon boundaries.

Table III: Chromosomal Mapping of the Calpactin I Heavy Chain Gene in Somatic Cell Hybrids[a]

hybrid	mouse chromosomes																				reaction with p36
BEM1-4	1,	2,	3,		5,	6,		8,			12,	13,	14,	15,	16,	17,	18,	19,	X		-
MACH4A63		2,					7,				12,	13,		15,	16,	17,	18,	19			-
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		+
MACH2A2C2	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							-

[a] Chromosomal contents of somatic cell hybrids were determined by karyotypic and isoenzyme analysis, and presence or absence of mouse DNA sequences reactive with the 3′ end p36 cDNA probe was determined by Southern blotting, as described (Saris et al., 1986). A chromosome is tallied as present if found at a frequency ≥0.15 copies/cell.

Table IV: Calpactin I Heavy Chain DNA Polymorphism in Recombinant Inbred Strains of Mice[a]

	B	X	D	strain																					
	1	2	6	8	9	1	1	1	1		1	1	1	1	2	2	2	2	2	2		2	3	3	3
	1	2	6	8	9	1	2	3	4		5	6	8	9	2	3	4	5	7	8		9	0	1	2
Call h	D	D	B	D	D	B	D	B	D		B	D	D	D	D	D	B	D	B	B		D	D	B	D
	B	X	H	strain														C	X	B	strain				
								1	1	1		1	1												
	2	3	4	6	7	8	9	0	1	2		4	9					D	E	G	H	I	J	K	
Call h	H	H	H	H	H	B	H	H	H	H		B	H					C	C	C	B	B	B	C	

[a] Recombinant inbred strains of mice (Taylor, 1981) were typed for the inheritance of the restriction fragment length polymorphism associated with the *Call h* locus by Southern blotting. All strains were homozygous for one of the progenitor strain alleles of the locus, as indicated by the single-letter codes: B, C57BL/6J-like; C, BALB/cJ-like; D, DBA/2J-like; H, C3H/HeJ-like.

coding p36 (Saris et al., 1986). To determine the chromosomal location of the gene, a panel of somatic hybrid cell lines carrying various combinations of mouse chromosomes on a constant Chinese hamster background was scored for the retention of mouse-specific DNA fragments cross-reactive with the cDNA probe. Three hybrids showed such fragments and six lacked them, although hamster-derived DNA fragments cross-reactive with the probe were visible in all cases. The only mouse chromosomes present in the positive hybrids but absent from the negative ones were 9 and 10 (Table III).

To distinguish between chromosomes 9 and 10, and to localize the p36 gene more precisely in a linkage group, we searched for a restriction fragment length polymorphism associated with the gene. Digestion of DNA from C57BL/6J mice with *Msp*I yielded a strongly hybridizing fragment of 4.8 kb, whereas DNA of nine other strains examined, O20/A, AKR/J, C57L/J, SWR/J, DBA/2J, C3H/HeJ, SJL/J, BALB/cJ, and STS/A, yielded a 4.1-kb fragment. The inheritance of these two alleles was followed in three informative panels of recombinant inbred mice (Taylor, 1978, 1981) (Table IV). Comparison of the strain distribution pattern for the p36 restriction fragment length polymorphism and ones previously determined for other loci in these strain panels showed significant linkage between the p36 locus and *Apoa1* and *Ncam*, two markers of proximal chromosome 9 (42–44) (Table V). Thus, the p36 gene maps to chromosome 9. We have named the locus *Call h* (calpactin I heavy chain).

Although the typing data from recombinant inbred strains were sufficient to establish linkage between *Call h* and other markers of mouse chromosome 9, they did not by themselves allow construction of an ordered linkage map because the number of informative strains was relatively small. We therefore used a map-building strategy applied previously to linkage groups on mouse chromosomes 11 (Bernier et al., 1988) and 12 (Blank et al., 1988). Data from published backcross experiments were used to construct a map of the linkage group at high confidence. Four loci previously placed in this map and also typed in recombinant inbred strains were then used as anchor loci in the construction of a map using recombinant inbred typing data. *Call h* was localized to the 15 cM (centimorgan) interval between *Apoa1* and *d* with a probability greater than 0.99 given the placement of the other markers. The best map of the region of chromosome 9 including *Call h* is (*Lap-1*)-7cM-(*Apoa1*,*Ncam*)-7cM-(*Call h*)-8cM-(*d*-)8cM-(*Mod-1*). *Apoa1* and *Ncam* are 3 cM apart but cannot be unambiguously ordered on the basis of available data.

Table V: Linkage of *Call h* to Other Markers Typed in Recombinant Inbred Mice[a]

locus		*D/I*	probability
Apoa1	(9)	7/41	0.00321
Ncam	(9)	8/42	0.00858
Mod-1	(9)	10/41	0.13842
Env-2	(9)	4/23	0.15976
Ltw-1	(9)	0/7	0.31133
Lap-1	(9)	11/40	0.39309
Mtv-13	(4)	5/23	0.41998
d	(9)	5/23	0.41998

[a] The strain distribution patterns obtained for the *Call h* locus were compared with those determined previously for 301 other loci that are informatively polymorphic in three strains of mice, and the Bayesian probability that the observed proportion of animals showing discordant inheritance of the two markers was a chance event was computed (D'Eustachio et al., 1985; Blank et al., 1988). All concordances associated with probabilities less than 0.5 are shown. The number in parentheses after each locus is the chromosome to which it is mapped. *D/I*, number of discordant/total number of informative recombinant inbred strains.

Discussion

At the protein level p36 is comprised of at least three functional domains, an N-terminal region of approximately 12 amino acids that interacts with p11 (Johnsson et al., 1988), a protease-sensitive linker region containing two phosphorylation sites at positions 23 (Tyr) and 25 (Ser) (Glenney &

Tack, 1985; Gould et al., 1986), and a C-terminal region, composed of four 75 amino acid tandem repeats (Kristensen et al., 1986; Saris et al., 1986), which has Ca^{2+}/phospholipid binding activity. The exact function of these four repeats remains unknown, but it is likely thay they are involved in the Ca^{2+}-dependent interaction of the protein with plasma membrane phospholipid. The exon/exon junctions within this part of the p36 gene do not correspond to the boundaries of the repeats found in p36 at the protein level (Figure 3). Furthermore, the 17 amino acid consensus sequence found by Geisow et al. (1986), which is present in multiple copies in all the members of the p36 family, is also not strictly delineated by intron/exon junctions, although in three cases there is an exon/exon boundary close to its N-terminal end (see below). It has been postulated that exons represent genetic building blocks that code for discrete structural or functional domains of the protein (Gilbert, 1978; Smith et al., 1979; Craik et al., 1982; Campbell & Porter, 1983). While not universally tenable, this hypothesis holds for the immunoglobulin, hemoglobin, and lysosyme gene families and perhaps also for some complement proteins. Either p36 does not conform to this hypothesis or the protein has more complex structure/function relationships than can be discerned from its primary structure.

DIAGON analysis of p36 cDNA sequence shows close homologies (>50%) within some exons (data not shown). Exons 3–11 share differing degrees of sequence homology, with the greatest homology found between exons 3, 5, 8, and 11, which implies that these exons arose by a duplicative mechanism. These exons each contain the 17 amino acid consensus sequence found by Geisow et al. (1986) and are thought to contain the Ca^{2+}/phospholipid binding sites. The relative position of each 17 amino acid consensus sequence within the exon, however, is not well conserved. In this context it is interesting to note that the exon/exon junctions within the Ca^{2+}-binding repeats of calmodulin also lack consistent positions (Simmen et al., 1985), even though this gene is believed to have evolved by duplication of a single Ca^{2+}-binding site. Although exon 12 does not have significant nucleotide homology with the equivalent exons in the other repeats, there is amino acid sequence similarity in exon 12 with other repeats (Figure 3). We do not know if exon 12 corresponds to any functional domain in p36, but it encodes the C terminus of the protein.

Although the intron/exon structure of the p36 gene does not fit well with the hypothesis that it arose by the duplication of a 75 amino acid repeat followed by a reduplication giving a four-repeat structure [see review by Crompton et al. (1988)], one structural feature is consistent with this idea. There are exon/exon boundaries between the Lys and Gly residues that lie in equivalent positions at the start of the 17 amino acid consensus in both repeats 2 and 4 (Figure 3). This is what would be expected if a two-repeat protein had been duplicated, since the second repeat in this protein would have given rise to the second and fourth repeats. However, the downstream exon/exon junctions in repeats 2 and 4 are not in equivalent positions. Although all the repeats are encoded by two exons, the positions of the other boundaries are not conserved. For example, there is also an exon/exon boundary between Lys and Gly at the start of the 17 amino acid consensus in repeat 1, but this is shifted by two residues in the alignment (Figure 3) and therefore is not equivalent to the junctions in repeats 2 and 4. If, as seems likely, the four repeats arose by two successive duplication events in an archetypal protein of this type, the exon/exon boundaries must subsequently have changed. In this regard it will be important to compare the exon organization of p36 with that of other members of this family. The exon structure within the conserved repeat regions of the other family members will be particularly revealing, because this may give us clues to the evolution of these proteins and insights into the mechanism of duplication.

In contrast to the situation for the C-terminal core of p36, the functional domains at the N terminus of p36 do correspond to individual exons. Exons 1 and 2 do not have significant nucleotide homology with the other exons; they code for amino acids 1–15 and 16–48, respectively. The site of p11 binding lies within the first 12 amino acids, the phosphorylation sites are found at positions 23 and 25, and the chymotrypsin-sensitive cleavage sites reside at positions 23 and 29. Thus, the p11-binding site lies completely within exon 1, whereas the two phosphorylation sites and the chymotrypsin cleavage sites lie within exon 2. A comparison of the exon organization of p36 with that of other members of this family will also be revealing here. The N-terminal region of each protein is unique, and one would expect that these regions would also be encoded by separate exons as is the case for p36.

An interesting feature of the p36 gene is the "exon clustering" (Figure 2); the first four exons are contained within roughly 10 kb and the last eight exons within 8 kb. These two groups of exons are separated by a 3.3-kb intron. It would be of interest to determine if this exon clustering has any functional relevance at the protein level.

The murine p36 cDNA clone analyzed by Saris et al. (1986) ends 7 nucleotides upstream of the ATG start codon. Primer extension analysis showed that there are ~52 nucleotides upstream of the end of the sequenced clone (Saris et al., 1986). We have recently cloned and sequenced a cDNA for the 5′ end of the p36 mRNA (C. J. M. Saris, T. Hunter, and B. F. Tack, unpublished data) and find this sequence to diverge from the genomic sequence at the CAGC lying between the two oligo-T runs (Figure 1A). We conclude that this CAGC must be used as a splice acceptor and that there must be at least one additional 5′ exon in the p36 gene. It should be noted that the sequence in this region of the genome is consistent with this being splice acceptor, since it contains the conserved AG and a pyrimidine-rich region upstream of the splice acceptor commonly found at the 3′ end of introns (Birnstiel et al., 1985). The 38-nucleotide cDNA sequence to the 5′ side of this splice site is homologous to the sequences of the 5′ ends of human and bovine p36 cDNAs (Huang et al., 1986; Kristensen et al., 1986). We are currently using this sequence to search for the p36 promoter region. No reactivity was detected with the RW22 cosmid clone, meaning that the promoter region must lie further than 20 kb upstream of the first coding exon.

We have mapped the mouse p36 gene (*Call*) to a locus on chromosome 9 approximately midway between *Ncam* and *d* (Table V). The calpactin light chain (p11) gene has previously been mapped to mouse chromosome 3 (Saris et al., 1987), and therefore the calpactin I heavy and light chains are unlinked. Preliminary results indicate that the mouse p35 gene maps to a site unlinked to either p36 or p11 (P. D'Eustachio, T. Hunter, and B. F. Tack, unpublished data). Analyses by Pepinsky et al. (1988) of rat calpactin/lipocortin proteins and by Huebner et al. (1988) of human calpactin/lipocortin genes indicate the existence of a large multigene family that, in humans, is widely dispersed over the genome. Huebner et al. (1988) identified four independently segregating loci with a human lipocortin II cDNA probe (the human p36 equivalent) and localized them to chromosomes 4, 9, 10, and 15. They also observed multiple mouse lipocortin II cross-hybridizing fragments, suggesting the possible presence of

multiple lipocortin II genes in the murine genome. We do not as yet understand the basis for the discrepancy in number of potential lipocortin II genes identified in the mouse. It seems unlikely that the results obtained by Huebner and co-workers can be due to cross-hybridization with other genes of the protein family, since there is no cross-hybridization between lipocortin II cDNA and lipocortin I cDNA. At the genetic level, the extensive conservation of linkage groups between humans and mice (Nadeau & Reiner, 1988) allows the following correlation. *Call h* in the mouse lies between chromosomal regions corresponding unambiguously to segments of human chromosomes 11q and 15p. The assignment of one human lipocortin II gene to the latter region and none to the former (Devereux et al., 1984) suggests that the 15p human locus encodes the human homologue of the murine calpactin I heavy chain gene mapped by our group. A combination of genetic and functional analyses of calpactin proteins and recombinant DNA clones in humans, rodents, and bovids will eventually enable us to unravel this large and fascinating gene family by enumerating its members, distinguishing among them structurally, and assigning a function to each.

In conclusion, we have established that the coding sequence of the p36 gene contains 12 exons and spans a region of at least 22 kb, that each of the two first exons encode discrete functional domains of the protein, and that the murine p36 structural gene is located on chromosome 9.

ACKNOWLEDGMENTS

We thank Deborah Noack and Lori Hicks-Graham for their excellent technical assistance and Bonnie Towle for typing the manuscript.

REFERENCES

Alitalo, K., Ralston, R., & Keski-Oja, J. (1984) *Exp. Cell Res. 150*, 177–185.

Bankier, A. T., & Barrell, B. G. (1983) in *Techniques in Nucleic Acid Biochemistry* (Flavell, R. A., Ed.) Vol. 85-08, pp 1–34, Elsevier/North-Holland Scientific, Limerick, Ireland.

Bernier, L., Colman, D. R., & D'Eustachio, P. (1988) *J. Neurosci. Res. 20*, 497–504.

Biggin, M. D., Gibson, T. J., & Hong, G. F. (1983) *Proc. Natl. Acad. Sci. U.S.A. 80*, 3963–3965.

Birnstiel, M. L., Busslinger, M., & Strub, K. (1985) *Cell 41*, 349–359.

Blank, R. D., Campbell, G. R., Calabro, A., & D'Eustachio, P. (1988) *Genetics 120*, 1073–1083.

Campbell, R. D., & Porter, R. (1983) *Proc. Natl. Acad. Sci. U.S.A. 80*, 4464–4468.

Chaplin, D. D., Woods, D. E., Whitehead, A. S., Goldberger, G., Colten, H. R., & Seidman, J. G. (1983) *Proc. Natl. Acad. Sci. U.S.A. 80*, 6947–6951.

Cheng, Y. S.-E., & Chen, L. B. (1981) *Proc. Natl. Acad. Sci. U.S.A. 78*, 2388–2392.

Cooper, J. A., & Hunter, T. (1981a) *Mol. Cell. Biol. 1*, 165–178.

Cooper, J. A., & Hunter, T. (1981b) *Mol. Cell. Biol. 1*, 394–407.

Cooper, J. A., & Hunter, T. (1983a) *J. Biol. Chem. 258*, 1108–1115.

Cooper, J. A., & Hunter, T. (1983b) *Curr. Top. Microbiol. Immunol. 107*, 125–161.

Courtneidge, S., Ralston, R., Alitalo, K., & Bishop, J. M. (1983) *Mol. Cell. Biol. 3*, 340–350.

Craik, C. S., Sprang, S., Fletterick, R., & Rutter, W. J. (1982) *Nature 299*, 180–182.

Crompton, M. R., Moss, S. E., & Crumpton, M. J. (1988) *Cell 55*, 1–3.

Decker, S. J. (1982) *Biochem. Biophys. Res. Commun. 109*, 434–441.

D'Eustachio, P., Owens, G. C., Edelman, G. M., & Cunningham, B. A. (1985) *Proc. Natl. Acad. Sci. U.S.A. 82*, 7631–7635.

Devereux, J., Haeberli, P., & Smithies, O. (1984) *Nucleic Acids Res. 12*(1), 387–395.

Drust, D. S., & Creutz, C. E. (1988) *Nature 331*, 88–91.

Erikson, E., & Erikson, R. L. (1980) *Cell 21*, 829–836.

Geisow, M. J., Fritsche, U., Hexham, J. M., Dash, B., & Johnson, T. (1986) *Nature 320*, 636–638.

Gerke, V., & Weber, K. (1984) *EMBO J. 3*, 227–233.

Gilbert, W. (1978) *Nature 271*, 501.

Glenney, J. R., Jr. (1985) *FEBS Lett. 192*, 79–82.

Glenney, J. R., Jr. (1986a) *Proc. Natl. Acad. Sci. U.S.A. 83*, 4258–4262.

Glenney, J. R., Jr. (1986b) *J. Biol. Chem. 261*, 7247–7252.

Glenney, J. R., Jr., & Tack, B. F. (1985) *Proc. Natl. Acad. Sci. U.S.A. 82*, 7884–7888.

Glenney, J. R., Jr., & Zokas, L. (1988) *Biochemistry 27*, 2069–2076.

Glenney, J. R., Jr., Tack, B., & Powell, M. A. (1987) *J. Cell Biol. 107*, 503–511.

Gould, K. L., Cooper, J. A., & Hunter, T. (1984) *J. Cell Biol. 98*, 487–497.

Gould, K. L., Woodgett, J. R., Isacke, C. M., & Hunter, T. (1986) *Mol. Cell Biol. 6*, 2738–2744.

Greenberg, M. E., & Edelman, G. M. (1983a) *Cell 33*, 767–779.

Greenberg, M. E., & Edelman, G. M. (1983b) *J. Biol. Chem. 258*, 8497–8502.

Grosveld, F. G., Lund, T., Murray, E. J., Mellor, A. L., Pahl, H. M., & Flavell, R. A. (1982) *Nucleic Acids Res. 10*, 6715–6732.

Huang, K.-S., Wallner, B. P., Mattaliano, R. J., Tizard, R., Burne, C., Frey, A., Hession, C., McGray, P., Sinclair, L. K., Chow, E. P., Browning, J. L., Ramachandran, K. L., Tang, J., Smart, J. E., & Pepinsky, R. B. (1986) *Cell 46*, 191–199.

Huebner, K., Cannizzaro, L. A., Frey, A. Z., Hecht, B. K., Hecht, F., Croce, C. M., & Wallner, B. P. (1988) *Oncogene Res. 2*, 299–310.

Hunter, T. (1988) *Adv. Exp. Med. Biol. 234*, 169–193.

Johnsson, N., Vandekerchove, J., Van Damme, J., & Weber, K. (1986) *FEBS Lett. 198*, 361–364.

Johnsson, N., Marriott, G., & Weber, K. (1988) *EMBO J. 7*, 2435–2442.

Kristensen, T., Saris, C. J. M., Hunter, T., Hicks, L., Noonan, D. J., Glenney, J. R., Jr., & Tack, B. F. (1986) *Biochemistry 25*, 4497–4503.

Lehto, V.-P., Virtanen, I., Ralston, R., & Alitalo, K. (1983) *EMBO J. 2*, 1701–1705.

Lusis, A. J., Taylor, B. A., Wangenstein, R. W., & LeBoeuf, R. C. (1983) *J. Biol. Chem. 258*, 5071–5078.

Mangeat, P. H., & Burridge, K. (1984) *J. Cell Biol. 98*, 1363–1377.

Maniatis, T., Fritsch, E. F., & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual*, Cold Spring Harber Laboratory, Cold Spring Harbor, NY.

Nadeau, J. H., & Reiner, A. H. (1988) *Curr. Top. Microbiol. Immunol. 137*, 39–40.

Nadeau, J. H., Kompf, J., Siebert, G., & Taylor, B. A. (1981) *Biochem. Genet. 19*, 465–474.

Nakamura, K. D., & Weber, M. J. (1982) *Mol. Cell. Biol. 2*, 147-153.
Nigg, E. A., Cooper, J. A., & Hunter, T. (1983) *J. Cell Biol. 96*, 1601-1609.
Pepinsky, R. B., Tizard, R., Mattaliano, R. J., Sinclair, L. K., Miller, G. T., Browning, J. L., Chow, E. P., Burne, C., Huang, K.-S., Pratt, D., Wachter, L., Hession, C., Frey, A. Z., & Wallner, B. P. (1988) *J. Biol. Chem. 263*, 10799-10811.
Radke, K., & Martin, G. S. (1979) *Proc. Natl. Acad. Sci. U.S.A. 76*, 5212-5216.
Radke, K., Gilmore, T., & Martin, G. S. (1980) *Cell 21*, 821-828.
Radke, K., Carter, V. C., Moss, P., Dehazya, P., Schliwa, M., & Martin, G. S. (1983) *J. Cell Biol. 97*, 1601-1611.
Sanger, F., Nicklen, S., & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U.S.A. 74*, 5463-5467.
Saris, C. J. M., Tack, B. F., Kristensen, T., Glenney, J. R., Jr., & Hunter, T. (1986) *Cell 46*, 201-212.
Saris, C. J. M., Kristensen, T., D'Eustachio, P., Hicks, L. J., Noonan, D. J., Hunter, T., & Tack, B. F. (1987) *J. Biol. Chem. 262*, 10663-10671.
Silver, J., & Buckler, C. E. (1986) *Proc. Natl. Acad. Sci. U.S.A. 83*, 1423-1427.
Simmen, R. C. M., Tanaka, T., Ts'ui, K. F., Patkey, J. A., Scott, M. J., Lai, E. C., & Means, A. R. (1985) *J. Biol. Chem. 260*, 907-912.
Smith, M., Leung, D. W., Gillam, S., & Astell, C. R. (1979) *Cell 16*, 753-761.
Staden, R. (1980) *Nucleic Acids Res. 8*, 3673-3694.
Staden, R. (1982) *Nucleic Acids Res. 10*, 4731-4751.
Taylor, B. A. (1978) in *Origins of Inbred Mice* (Morse, H. C., III, Ed.) pp 423-428, Academic Press, New York.
Taylor, B. A. (1981) in *Genetic Variants and Strains of the Laboratory Mouse* (Green, M. C., Ed.) pp 397-407, Gustav Fischer, Stuttgart, FRG.
Zokas, L., & Glenney, J. R., Jr. (1987) *J. Cell Biol. 105*, 2111-2121.

Organization of the Gene for Platelet Glycoprotein IIb†

Randy Heidenreich,‡ Robin Eisman,§ Saul Surrey,§,∥ Kathleen Delgrosso,§ Joel S. Bennett,⊥ Elias Schwartz,§,∥ and Mortimer Poncz*,§

Divisions of Hematology and Metabolic Diseases, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania 19104, Division of Hematology and Oncology, Hospital of the University of Pennsylvania, Philadelphia, Pennsylvania 19104, and Departments of Pediatrics, Human Genetics, and Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104

Received July 12, 1989; Revised Manuscript Received September 21, 1989

ABSTRACT: The glycoprotein (GP) IIb/IIIa heterodimer functions as a receptor for fibrinogen, von Willebrand factor, and fibronectin on activated platelets; it is dysfunctional in the bleeding diathesis Glanzmann's thrombasthenia. This receptor is a member of the integrin family, which includes homologous membrane receptors involved in a number of different cell-cell and cell-matrix adhesive interactions. Knowledge of the sequence and organization of the GPIIb and GPIIIa genes will help in understanding evolutionary relationships and functional homologies of this family of adhesion protein receptors and will facilitate analysis of molecular defects responsible for thrombasthenia. Using the GPIIb cDNA as a probe, we have isolated overlapping genomic clones encompassing the entire coding region, the 5′- and 3′-untranslated sequences, and the immediate flanking regions for the GPIIb gene. The gene spans approximately 17.2 kilobases (kb); all but approximately 2.6 kb of intronic DNA sequence has been determined. The GPIIb gene contains 30 exons whose demarcations do not correlate with previously suggested functional domains. Two intron/exon borders have the rare GC splice donor sequence instead of the consensus GT sequence. There are at least seven complete and three partial *Alu*I sequence repeats within the intron sequences. RNase protection, S1 nuclease analysis, and primer extension studies using human erythroleukemia (HEL) cell RNA and platelet RNA map a major transcription start site 32 base pairs (bp) 5′ to the beginning of the coding region; however, there are no canonical consensus TATA or CAAT boxes in the region immediately 5′ to the proposed cap site. The immediate 5′-flanking sequence of rodent GPIIb demonstrates complete identity near the proposed cap site with its human counterpart, but again, no TATA or CAAT boxes are apparent.

The GPIIb/IIIa complex, a calcium-dependent heterodimer present in platelet membranes, is the binding site for fibrinogen (Jennings & Phillips, 1982; Bennett et al., 1982, 1983), fibronectin, and von Willebrand factor (Ruggeri et al., 1982; Ginsberg et al., 1983) on activated platelets. Glycosylated GPIIb has a M_r of 136 000 on SDS-polyacrylamide gels and contains two disulfide-linked subunits: a heavier M_r 125 000 subunit and a lighter M_r 23 000 subunit (Jennings & Phillips, 1982). GPIIIa is a single-chain disulfide bond rich glycoprotein with an unreduced apparent M_r of 95 000 that increases to 110 000 following disulfide reduction (Jennings & Phillips, 1982). Defects or deficiencies of the GPIIb/IIIa heterodimer underlie the autosomal recessive bleeding disorder Glanzmann's thrombasthenia. GPIIb/IIIa is a member of the integrin family of membrane adhesion receptors that mediate cell-cell and cell-matrix interactions. This family includes

† Supported in part by grants from the National Institutes of Health (HL37419, AM16691, HL40387, HL28157, HD07107), the March of Dimes, and the Council for Tobacco Research-U.S.A., Inc.
* Author to whom reprint requests and correspondence should be addressed at The Children's Hospital of Philadelphia.
‡ Division of Metabolic Diseases, The Children's Hospital of Philadelphia.
§ Division of Hematology, The Children's Hospital of Philadelphia, and Department of Pediatrics, University of Pennsylvania.
∥ Department of Human Genetics, University of Pennsylvania.
⊥ Division of Hematology and Oncology, Hospital of the University of Pennsylvania, and Department of Medicine, University of Pennsylvania.

0006-2960/90/0429-1232$02.50/0