The human and murine protocadherin-β one-exon gene families show high evolutionary conservation, despite the difference in gene number

Koen Vanhalst[1], Patrick Kools[1], Eveline Vanden Eynde, Frans van Roy*

Molecular Cell Biology Unit, Department of Molecular Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, Ledeganckstraat 35, B-9000 Ghent, Belgium

Abstract Extensive cDNA analysis demonstrated that all human and mouse protocadherin-β genes are one-exon genes. The protein sequences of these genes are highly conserved, especially the three most membrane-proximal extracellular domains. Phylogenetic analysis suggested that this unique gene family evolved by duplication of one single protocadherin-β gene to 15 copies. The final difference in the number of protocadherin-β genes in man (#19) and mouse (#22) is probably caused by duplications later in evolution. The complex relationship between human and mouse genes and the lack of pseudogenes in the mouse protocadherin-β gene cluster suggest a species-specific evolutionary pressure for maintenance of numerous protocadherin-β genes. © 2001 Published by Elsevier Science B.V. on behalf of the Federation of European Biochemical Societies.

Key words: Protocadherin-β; Phylogeny; One-exon gene; Gene family

1. Introduction

Cadherins are calcium-dependent transmembrane (TM) glycoproteins belonging to a superfamily of over 100 members [1,2]. Evidence for the existence of six different cadherin subfamilies has been obtained by phylogenetic analysis [3]. All cadherins have in common that they contain an extracellular domain comprising five or more cadherin-specific repeats of approximately 110 amino acid residues. Although conservation between repeats of different cadherins is not extremely high, several small regions are well conserved and believed to be important in calcium-binding and correct folding of proteins [4–7]. Most cadherins contain a single TM domain and a cytoplasmic tail, differing throughout the several subfamilies.

One large subfamily comprises the protocadherins, which are characterized by containing more than five extracellular cadherin repeats (ECs) and by completely deviating cytoplasmic tails. This suggests interactions with other groups of cytoplasmic proteins, and probably also individual functions for the different subfamilies [8]. Three protocadherin subfamilies,

termed PCDHα, PCDHβ and PCDHγ, have been identified on human chromosome 5q31–q33 [9]. They all have a remarkable genomic organization: the EC and TM domains as well as part of the cytoplasmic domains are encoded by a single exon, organized in a tandem array. This is in contrast to the genomic structure of classical cadherins [1,3]. The remaining part of the cytoplasmic domains of PCDHα and PCDHγ is constant, though specific for the proteins of the two families. This carboxy-terminal sequence is encoded by three smaller exons located downstream of the cluster of tandemly arranged 5′ exons [9].

The human PCDHβ family has been reported to contain 16 genes and three pseudogenes, while the most recently published mouse Pcdhβ gene cluster comprises 22 genes [9,10]. Similar results were previously obtained in rat, for which at least 10 variants were suggested [11]. For two of the latter variants in situ hybridization showed a differential expression pattern in various regions of the mouse brain [11]. Transfection studies with Pcdh3 cDNA showed weak Ca^{2+}-dependent cell aggregation activity, sensitive to trypsin. This is in contrast with the strong adhesion properties of most classical cadherins and suggests a distinctive biological role for protocadherin-β proteins.

Previous publications predicted that members of the protocadherin-β family would also use common exons at the 3′ end of the transcripts, as demonstrated for the two related protocadherin-α and -γ families [9]. Here we report on the further investigation of this particular assumption. In addition, both human and murine protocadherin-β gene clusters were reconstructed and compared through phylogenetic analysis.

2. Materials and methods

2.1. Identification and isolation of cDNAs

Expressed sequence tag (EST) libraries (dbEST, NCBI) covering mouse and man were screened with BLAST-N [12], using sequence data from the rat *Pcdh3* gene (GenBank accession number L43592). EST clones were obtained from the UK-HGMP Resource Center (Hinxton, UK) or Genome Systems (St. Louis, MO, USA), and were completely sequenced. To obtain additional human cDNA clones, a human fetal brain 5′ stretch λDR2 phage library (Clontech, Palo Alto, CA, USA) was screened using the insert of human EST clone IMAGE 30294. To isolate murine cDNA sequences, a 12-dpc mouse embryo cDNA library (Novagen, Madison, WI, USA) in the λEXlox vector was screened using the insert of EST clone IMAGE 789667.

2.2. Completion of cDNA sequences

RACE (rapid amplification of cDNA ends) experiments were performed according to the Marathon cDNA amplification protocol (Clontech), using human fetal brain RNA (Clontech) after poly(A)+

---

*Corresponding author. Fax: (32)-9-2645331 or 2645348.
E-mail: f.vanroy@dmb.rug.ac.be

[1] Both authors contributed equally.

selection (Micro-FastTrack; Invitrogen, San Diego, CA, USA) or Marathon-ready mouse brain cDNA (Clontech). RACE PCR reactions were performed using the Advantage cDNA polymerase mix (Clontech). Primers used for 3′ RACE on mouse brain cDNA were MCB1216 (5′-CACTCATCTCCATCAACGCTGAC-3′) as gene-specific primer GSP1 and MCB1217 (5′-CATAGTGGTGCTGGACGA-CAATG-3′) as GSP2. For 3′ RACE experiments on human fetal brain cDNA, primers MCB1204 (GSP1) (5′-CTTCTCCCAGCCC-TACCTGCCT-3′) and MCB1205 (GSP2) (5′-CCAGGGCATCTGG-TGGACGTGAG-3′) were used. Reverse primers were AP1 and AP2 as supplied within the Marathon kit (Clontech). Sequence analysis was performed using the dideoxy chain termination method with fluorescent detection on an ABI Prism 377 apparatus (Perkin-Elmer, Foster City, CA, USA) with various specific primers, designed using Oligo v5.0 software (Medprobe, Plymouth, MN, USA) and synthesized by Life Technologies (N.V. Life Technologies SA, Merelbeke, Belgium).

2.3. In silico analysis

The human gene cluster was reconstructed using BAC sequences in the public domain (GenBank accession numbers AC008688, AC05754 and AC05752) and STADEN software [13]. The gap between the last two BAC sequences was closed using GenBank accession number AF282973. The murine gene cluster was reconstructed by sequencing selected cosmids and on the basis of draft sequences of murine BACs (GenBank accession numbers AC020967, AC020973, AC020974 and AC069448). Protein sequences were obtained by in silico translation of open reading frames (ORFs). Subsequent multiple alignment and phylogenetic analysis were performed using the CLUSTAL-X program [14]. Bootstrap neighbor joining trees were drawn with the TREEVIEW program [15]. Putative signal peptide sequences and TM domains were predicted with dedicated software (http://www.cbs.dtu.dk/services). Box shading was done with the WWW-BOXSHADE server (http://www.ch.embnet.org/software/BOX_form.html).

3. Results and discussion

3.1. PCDHβ genes are one-exon genes

It was suggested that the PCDHβ family members contain additional 3′ exons encoding a cytoplasmic tail, as was shown for the PCDHα and PCDHγ family members [9]. The presence of highly conserved putative 5′ splice sites at the end of most protocadherin-β ORFs in both man and mouse further emphasized the possible presence of additional 3′ exons [10]. Extensive comparison of all public domain EST sequences as well as 26 human and five murine cDNA clones isolated and sequenced by us did however not reveal any additional 3′

Table 1
Overview of the human (top) and murine (bottom) protocadherin-β cDNA clones analyzed

Alias	Gene symbol[a]	cDNA sequence source	Poly(A) signal	PXXP domain	GenBank accession number
β1	PCDHβ1	–	?	Yes	AF152488
β2	PCDHβ2	EST (345262)+cDNA (pGD27)	Yes	Yes	AF217756
β3	PCDHβ3	cDNA (pGD40, pGD45)+3′ RACE	Yes	Yes	AF217755
β4	PCDHβ4	EST (363288, 380379)+cDNA (pGD33)+3′ RACE	Yes	Yes	AF217754
β5	PCDHβ5	EST (25498, 729368, 729372)	Yes	Yes	AF217753
β6	PCDHβ6	3′ RACE	Yes	Two	AF217752
ψ1	PCDHβ17P	cDNA (pGD3, pGD8, pGD9, pGD16)+3′ RACE	Yes	No	AF217751
β7	PCDHβ7	cDNA (pGD1, pGD5, pGD14, pGD17, pGD18, pGD41)	Yes	Yes	AF217750
β8	PCDHβ8	cDNA (pGD12, pGD42, pGD43)	Yes	Yes	AF152501
β16	PCDHβ16	EST (1840729, 1840753)	No	Yes	AF217757
β9	PCDHβ9	cDNA (pGD11, pGD25)	No	Two	AF217749
β10	PCDHβ10	EST (25205)+5′ RACE+3′ RACE	Yes	No	AF217748
β11	PCDHβ11	cDNA (pGD34, pGD35)+5′ RACE+3′ RACE	Yes	Yes	AF217747
β12	PCDHβ12	3′ RACE	Yes	Yes	AF217746
β13	PCDHβ13	cDNA (pGD7)	No	Yes	AF217745
β14	PCDHβ14	cDNA (pGD15)	Yes	Yes	AF217744
ψ2	PCDHβ18P	cDNA (pGD4)+3′ RACE	No	Yes	AF217743
ψ5	PCDHβ19P	cDNA (pGD24, pGD32)	No	No	AF329369
β15	PCDHβ15	EST (30294)	Yes	Yes	AF217742
A	Pcdhβ1	–	?	Yes	AF326294
B	Pcdhβ2	–	?	Yes	AF326295
C	Pcdhβ3	cDNA (pKV3.1) EST (789667)	Yes	Yes	AF326296
D	Pcdhβ5A	–	?	Yes	AF326297
E	Pcdhβ4A	–	?	Yes	AF326298
F	Pcdhβ5B	–	?	No	AF326299
G	Pcdhβ4B	cDNA (pKV1.1, pKV3.2)+3′ RACE	Yes	Yes	AF326300
H	Pcdhβ5C	–	?	Yes	AF326301
I	Pcdhβ4C	–	?	Yes	AF326302
J	Pcdhβ5D	–	?	No	AF326303
K	Pcdhβ5E	EST (1124384)	No	Yes	AF326304
L	Pcdhβ5F	–	?	Two	AF326305
M	Pcdhβ6	–	?	Yes	AF326306
N	Pcdhβ17	–	?	Yes	AF326307
O	Pcdhβ7	EST (1122311)+3′ RACE	No	No	AF326308
P	Pcdhβ8	–	?	Yes	AF32639
Q	Pcdhβ16	EST (1049529, 902840)+3′ RACE	Yes	Yes	AF326310
R	Pcdhβ9	–	?	No	AF326311
S	Pcdhβ11	cDNA (pKV1.9)+3′ RACE	Yes	Yes	AF326312
T	Pcdhβ14	EST (775497)	Yes	Yes	AF326313
U	Pcdhβ18	3′ RACE	Yes	No	AF326314
V	Pcdhβ15	EST (1381298, 1498571, 1499022, 1052992)+cDNA (pKV4.1)+3′ RACE	Yes	No	AF326315

[a]Assigned by the official human and mouse gene nomenclature commissions.

exons (Table 1). Subsequent 3′ RACE experiments with primers in the highly conserved sequence encoding the membrane-proximal extracellular domain (MPED) and TM region again did not lead to the identification of such exons, neither in mouse nor in man. All sequences of the 3′ RACE products were identical to the genomic DNA adjacent to the ORFs. The presence of 3′ exons is very unlikely, as they were not detected in over 400 cDNA clones tested. This strongly suggests that the protocadherin-β family comprises a large group of tightly clustered one-exon genes.

Several arguments indicate that additional 3′ exons could

have been present in the distant past. First, all three protocadherin families are closely related and two of them do contain constant exons. Second, the conserved parts of the ORFs of the PCDHβ family terminate at about the same position (i.e. at a PXXP motif; Fig. 1) as the first exon of the PCDHα and PCDHγ families do. We hypothesize that in the course of evolution, more particularly before separation of the mouse and human mammalian lines, the PCDHβ family members lost these exons. This has either been advantageous or tolerated, as this last constant part of the proteins might have had only limited functionality in the protocadherin-β ancestor.
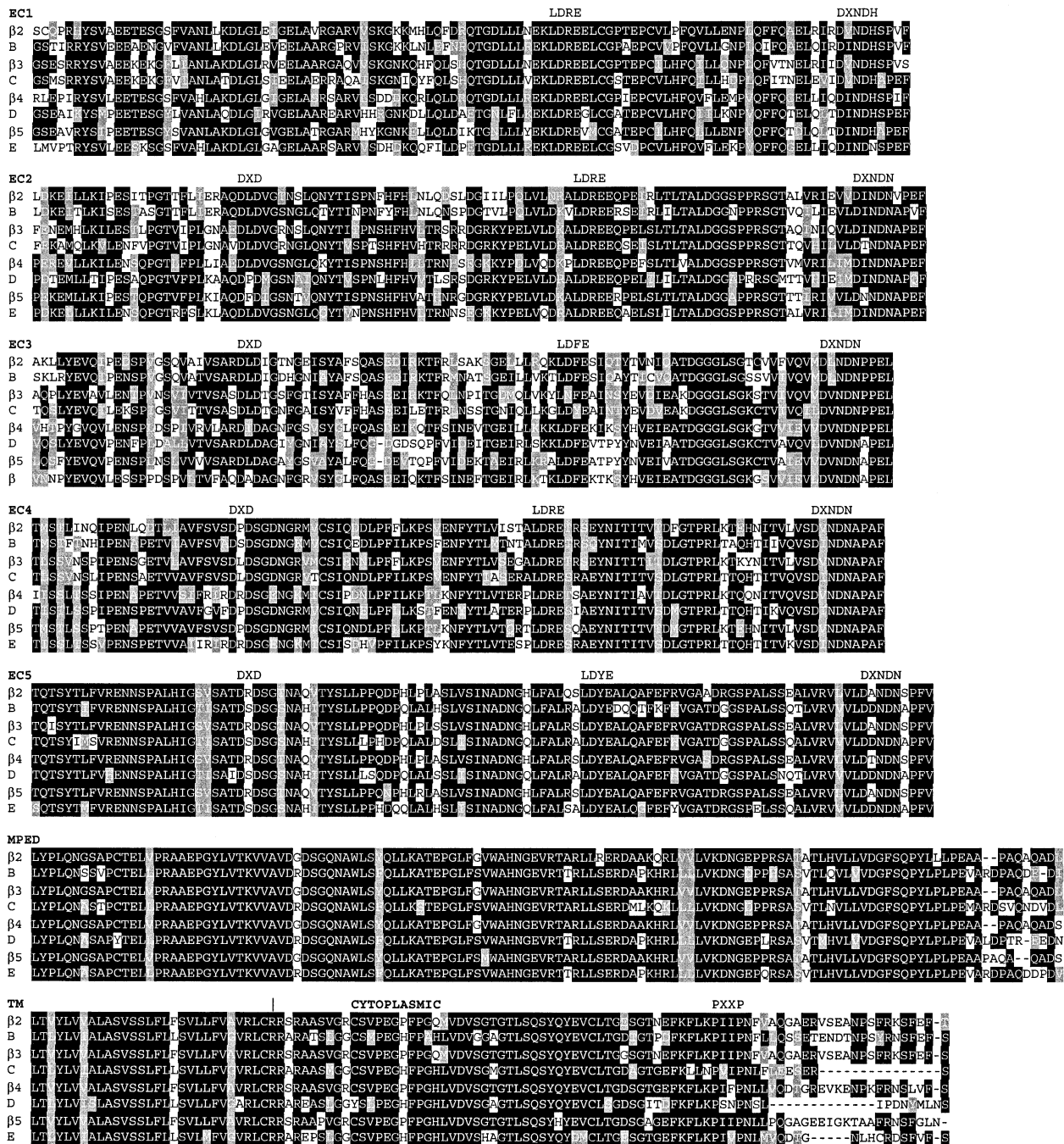
Fig. 1. Alignment of five representative human and murine protein sequences starting behind the putative signal peptide. Protein domains (EC1–EC5, MPED, TM, cytoplasmic) are aligned separately. Conserved cadherin motifs are indicated above the alignments. Black background indicates identities of more than 50%. Gray shading implicates amino acid residues identical in more than 30%, but less than 50% of the cases.
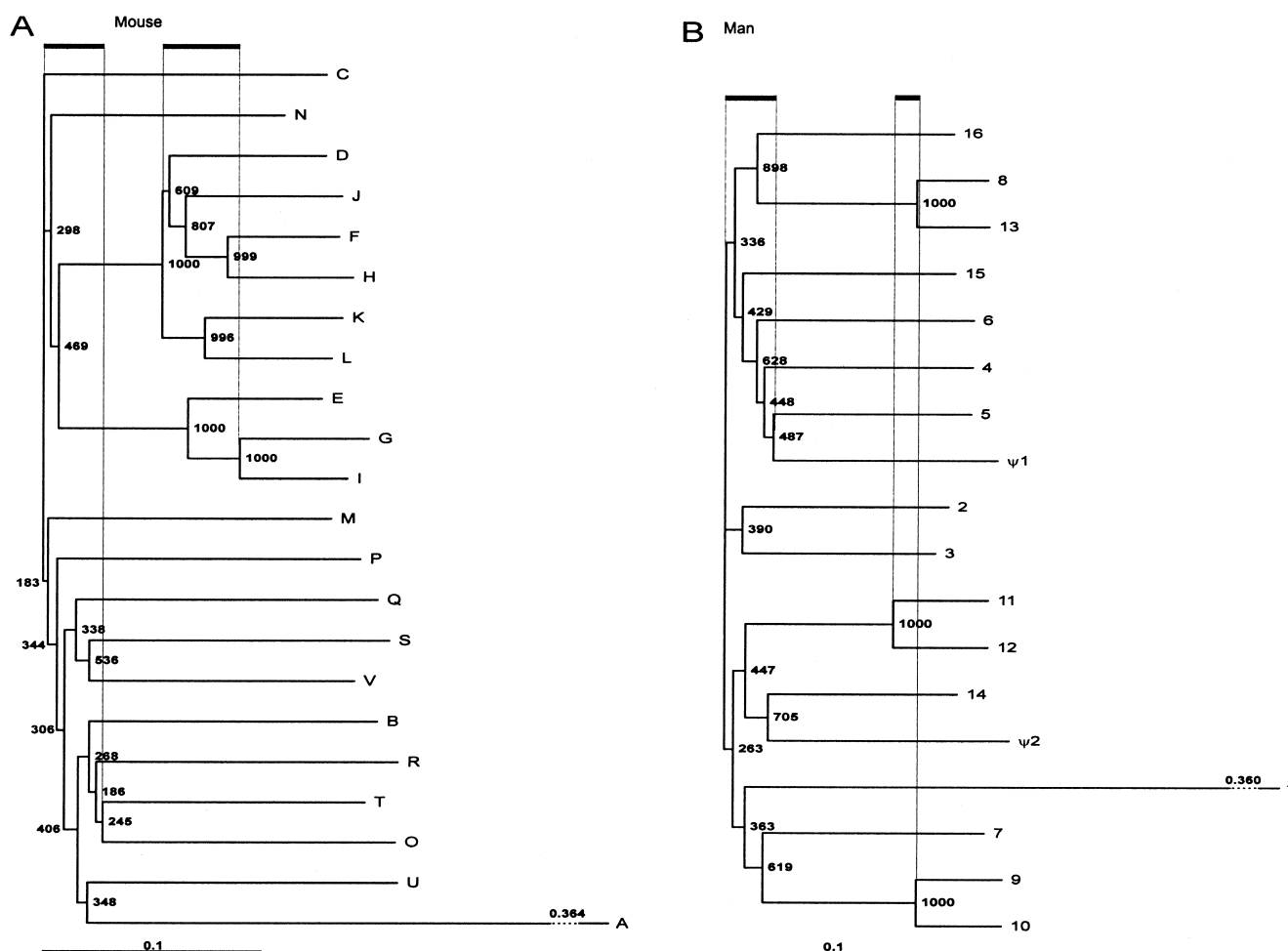
Fig. 2. Rooted tree view of CLUSTAL-X alignments of murine (A) and human (B) protocadherin-β protein sequences. Scale bar corresponds to one amino acid substitution per 10 amino acid residues. Two clusters of divergence points (branching) are indicated by horizontal bars on top of both murine and human trees. Divergence lines for the murine PcdhβA gene and human PCDHβ1 gene were shortened, their original length being indicated.

3.2. Comparison of predicted protein sequences

On the basis of the human and mouse genomic sequences accessible in the public domain or assembled by us, we concluded that the human PCDHβ cluster comprises 16 genes and three pseudogenes (PCDHβ1 through PCDHβ16, ψ1, ψ2 and ψ5) whereas the mouse Pcdhβ cluster comprises 22 genes (termed PcdhβA through PcdhβV). This is in full agreement with the recently published data of Wu et al. [10]. All predictable full-length human PCDHβ and murine Pcdhβ protein sequences were aligned, which revealed high mutual similarities (illustrated in Fig. 1 for five human and five murine protein sequences). Interestingly, the level of similarity between the various EC domains appears to be divergent: EC1, EC2 and EC3 show the lowest level of similarity, whereas EC4, EC5 and EC6 (MPED) are highly conserved. An exception to this is the similarity between the β8 and β13 proteins, for which EC1 is completely identical while the other domains are highly similar (not shown). The low overall conservation between the first three EC repeats of the other proteins favors the idea of specific, homophilic adhesion phenomena. Increasing evidence indicates that not only the EC1 repeat, but many of the other EC repeats are involved in establishing firm cell–cell contacts [16]. Since EC4, EC5 and EC6 are highly conserved, these domains might be of impor-

tance for strengthening the adhesive properties, this by lateral *cis*-dimerization or even by interaction with other proteins.

The single TM and cytoplasmic domains are highly conserved among all PCDHβ family members, except for the last 20 amino acid residues. This difference in carboxy-terminal tail is likely due to the evolutionary loss of 3′ exons as discussed above. At the very end of the conserved region of the cytoplasmic domain, most protein sequences reveal a PXXP amino acid motif (Fig. 1), being a putative SH3 protein-binding site [17]. In man, only PCDHβ10 lacks this site, while six out of the 22 murine proteins do not contain this putative domain either. However, some human and murine protocadherin-β proteins even have two such domains next to each other. In contrast, the four SH3 sites in each PCDHα (cadherin-related neuronal receptor) family member are encoded by the constant 3′ exons [18]. The presence of putative SH3-binding domains in many PCDHβ proteins suggests that the loss of the constant 3′ exons might not have completely abolished the original function of an elongated cytoplasmic tail.

3.3. Phylogenetic comparison of murine and human PCDHβ gene clusters

To perform a correct phylogenetic analysis, sequences of

approximately the same length should be used. In the case of the murine gene cluster, all coding sequences, except for the first gene PcdhβA, had about the same length. However, the two previously reported human PCDHβ pseudogenes [9] have in-frame mutations leading to either a carboxy-terminally or an amino-terminally truncated protein. For the purpose of the alignment, we repaired these mutations in silico, generating full-length sequences highly similar to the other PCDHβ genes. Furthermore, we only used protein sequences beginning after the signal peptide (EC1) and ending at the PXXP motif of the cytoplasmic tail.

First, the human and murine protein sequences were separately subjected to a phylogenetic analysis. This yielded similar results in both species (Fig. 2A,B). Phylogenetic analysis of isolated protein domains (data not shown) was largely in line with these results.

Interestingly, several genes appear to be recently diverged from a single ancestor gene and can be considered paralogs, e.g. β8/β13, β9/β10, and β11/β12 in man, and the genes D/F/H/J/K/L, and E/G/I in mouse. The other genes appear to have deviated relatively early in evolution (Fig. 2A,B). Moreover, the branch nodes in the human and murine phylogenetic trees, indicated by bars on top of Fig. 2A,B, seem to cluster in two groups, most likely corresponding to two periods of gene duplication during evolution. The phylogenetic trees indicate that both human and murine mammalian ancestors comprised 15 protocadherin-β 'precursor' genes at the end of the suggested first period. Subsequent gene duplications in the second period led to the differences in gene number between the human and murine gene cluster. The first wave of duplications is expected to have occurred relatively late in evolution, as no

relatives of the protocadherin-β genes were identified in lower animals like *Caenorhabditis elegans* and *Drosophila melanogaster*. We propose that the second duplication period occurred after mammalian radiation, when human and murine lines were already separated.

All human and murine protocadherin-β protein sequences were also compared in the same alignment (Fig. 3), showing that several human and murine sequences were clustering together. Similar conclusions were drawn by Wu et al. [10]. Comparison of these clusters with the genomic position of the corresponding genes revealed an obvious conservation of the gene order in most cases (Fig. 4). The most striking examples are PCDHβ1/PcdhβA, PCDHβ2/PcdhβB and PCDHβ3/PcdhβC, therefore being most likely orthologs of each other. For the other molecules, the relationships were far more complex (Figs. 3 and 4). The murine paralogous genes PcdhβE, PcdhβG and PcdhβI most strongly resemble a single human gene PCDHβ4. The murine paralogous genes PcdhβD, PcdhβF, PcdhβH, PcdhβJ, PcdhβK and PcdhβL show the highest similarity to only one human gene, being PCDHβ5. In contrast, the human paralogous genes PCDHβ9 and PCDHβ10 match only one murine gene, the PcdhβR gene. The flanking murine gene PcdhβS shows the highest similarity to human PCDHβ11 and PCDHβ12. In all these instances, the genes involved are neighbors to each other. In one case only, the closest family members PCDHβ8 and PCDHβ13 are separated by five other genes (Fig. 4). Strikingly, the overall similarity between these two proteins is 92.1%, the highest value observed (data not shown). The relation between the mouse PcdhβV gene and human ψ5 is based on nucleotide comparisons (data not shown). Also striking is the low overall
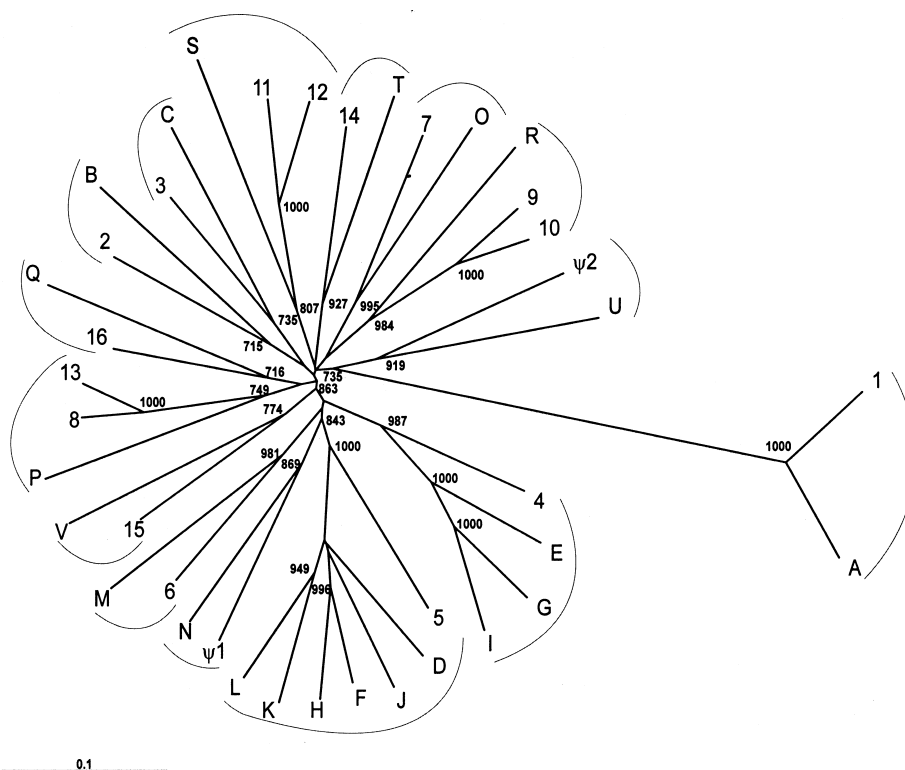

Fig. 3. Unrooted tree of a CLUSTAL-X alignment of murine (A through V) and human (β1 through β16, ψ1 and ψ2) protein sequences. Only bootstrap values above 700 are given as they represent an accuracy of more than 95% for the predicted branches. Scale bar corresponds to one amino acid substitution per 10 amino acid residues.
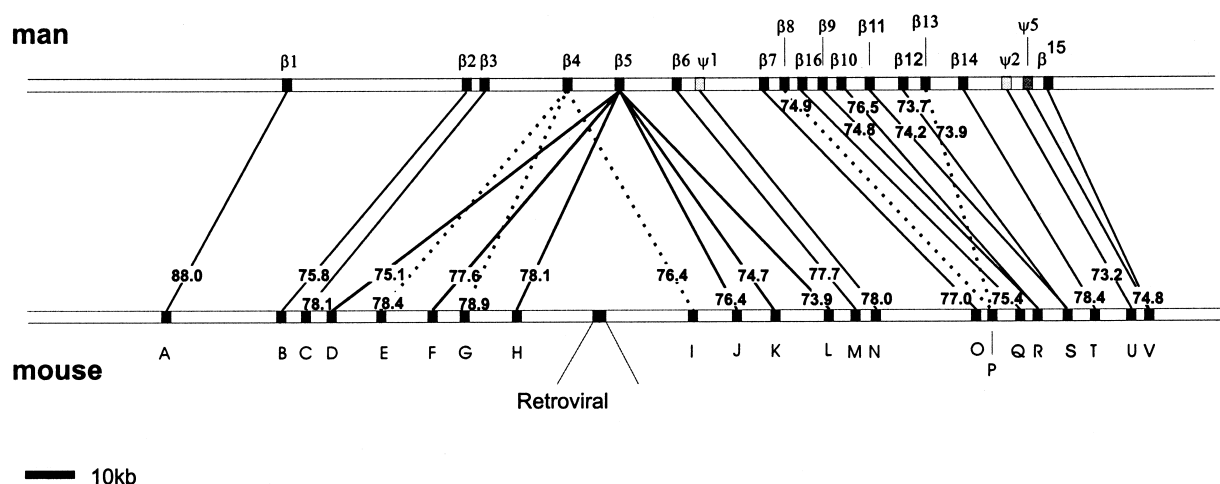
Fig. 4. Schematic representation of the human and murine protocadherin-β gene clusters. Horizontal lines represent the human and murine genomic sequences. ORFs of the one-exon genes are shown as black boxes, whereas gray boxes indicate pseudogenes. Connecting lines between ORFs of the two different clusters indicate genes with the highest similarity at the protein level (percentages of similarity are as indicated). For the human ψ5 gene, the similarity was determined at the nucleotide level. Scale bar, 10 kbp.

similarity of the human PCDHβ1 and murine PcdhβA sequences as compared to the other genes of the family, while their interspecies conservation is the highest (Figs. 2 and 3). Both in man and mouse, this is the first gene of the tandemly organized cluster (Fig. 4). This observation suggests that the orthologous PCDHβ1 and PcdhβA might be direct descendants of the ancestor gene.

It is clear from these data that the protocadherin-β genes have arisen by repeated gene duplication events. Such phenomena occur quite frequently in mammalian genomes and many of the duplicated genes subsequently become non-functional pseudogenes or are deleted from the genome. However, this is not the case here suggesting an increasing functional need for a large number of similar genes. The mechanism of obtaining additional genes has been called the birth-and-death process of evolution and is responsible for the formation of many multigene families, such as the major histocompatibility and immunoglobulin gene families [19,20].

In conclusion, we reconstructed both the murine Pcdhβ and human PCDHβ gene clusters and could demonstrate that they consist of closely packed one-exon genes. We subsequently compared the murine and human protocadherin-β gene clusters and showed a striking conservation, next to a remarkable difference in gene number. Regarding the function of these genes in either mouse or man, one can only speculate at this point. Protocadherin-β genes are expressed in murine brain in a region-specific manner, and the genes are located in a chromosomal region implicated in several neurological disorders [11]. Taken together with the weak cell–cell adhesion properties of protocadherins, this might point to a function in the development of the vertebrate brain. Functional analysis, including identification of molecular interaction partners and mouse knock-out studies will for sure widen our views on the role of the protocadherin-β family.

References

[1] Suzuki, S.T. (1996) J. Cell. Biochem. 61, 531–542.
[2] Yagi, T. and Takeichi, M. (2000) Gene Dev. 14, 1169–1180.
[3] Nollet, F., Kools, P. and van Roy, F. (2000) J. Mol. Biol. 299, 551–572.
[4] Blaschuk, O.W., Sullivan, R., David, S. and Pouliot, Y. (1990) Dev. Biol. 139, 227–229.
[5] Shapiro, L. et al. (1995) Nature (Lond.) 374, 327–337.
[6] Shapiro, L., Kwong, P.D., Fannon, A.M., Colman, D.R. and Hendrickson, W.A. (1995) Proc. Natl. Acad. Sci. USA 92, 6793–6797.
[7] Nagar, B., Overduin, M., Ikura, M. and Rini, J.M. (1996) Nature (Lond.) 380, 360–364.
[8] Suzuki, S.T. (1996) J. Cell Sci. 109, 2609–2611.
[9] Wu, Q. and Maniatis, T. (1999) Cell 97, 779–790.
[10] Wu, Q. et al. (2001) Genome Res. 11, 389–404.
[11] Sago, H. et al. (1995) Genomics 29, 631–640.
[12] Altschul, S.F., Warren, G., Miller, W., Myers, E.W. and Lipman, D.J. (1990) J. Mol. Biol. 215, 403–410.
[13] Bonfield, J.K., Smith, K.F. and Staden, R. (1995) Nucleic Acids Res. 23, 4992–4999.
[14] Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. and Higgins, D.G. (1997) Nucleic Acids Res. 24, 4876–4882.
[15] Page, R.D.M. (1996) Comput. Appl. Biosci. 12, 357–358.
[16] Sivasankar, S., Brieher, W., Lavrik, N., Gumbiner, B. and Leckband, D. (1999) Proc. Natl. Acad. Sci. USA 96, 11820–11824.
[17] Alexandropoulos, K., Cheng, G. and Baltimore, D. (1995) Proc. Natl. Acad. Sci. USA 92, 3110–3114.
[18] Sugino, H., Hamada, S., Yasuda, R., Tuji, A., Matsuda, Y., Fujita, M. and Yagi, T. (2000) Genomics 63, 75–87.
[19] Gaudieri, S., Kulski, J.K., Balmer, L., Giles, K.M., Inoko, H. and Dawkins, R.L. (1997) DNA Seq. 8, 137–141.
[20] Nei, M., Gu, X. and Sitnikova, T. (1997) Proc. Natl. Acad. Sci. USA 94, 7799–7806.