DNA RESEARCH 5, 355–364 (1998)

Prediction of the Coding Sequences of Unidentified Human Genes. XII. The Complete Sequences of 100 New cDNA Clones from Brain Which Code for Large Proteins *in vitro*

Takahiro NAGASE, Ken-ichi ISHIKAWA, Mikita SUYAMA, Reiko KIKUNO, Makoto HIROSAWA, Nobuyuki MIYAJIMA, Ayako TANAKA, Hirokazu KOTANI, Nobuo NOMURA, and Osamu OHARA*

Kazusa DNA Research Institute, 1532-3 Yana, Kisarazu, Chiba 292-0812, Japan

(Received 20 November 1998)

Abstract

In this paper, we report the sequences of 100 cDNA clones newly determined from a set of size-fractionated human brain cDNA libraries and predict the coding sequences of the corresponding genes, named KIAA0819 to KIAA0918. These cDNA clones were selected on the basis of their coding potentials of large proteins (50 kDa and more) by using *in vitro* transcription/translation assays. The sequence data showed that the average sizes of the inserts and corresponding open reading frames are 4.4 kb and 2.5 kb (831 amino acid residues), respectively. Homology and motif/domain searches against the public databases indicated that the predicted coding sequences of 83 genes were similar to those of known genes, 59% of which (49 genes) were categorized as coding for proteins functionally related to cell signaling/communication, cell structure/motility and nucleic acid management. The chromosomal locations and the expression profiles of all the genes were also examined. For 54 clones including brain-specific ones, the mRNA levels were further examined among 8 brain regions (amygdala, corpus callosum, cerebellum, caudate nucleus, hippocampus, substantia nigra, subthalamic nucleus, and thalamus), spinal cord, and fetal brain.

Key words: large proteins; *in vitro* transcription/translation; cDNA sequencing; expression profile; chromosomal location; brain

1. Introduction

The importance of human cDNA sequencing is widely accepted because it is expected to offer protein coding information more directly than the genomic sequencing. As the human genome project has entered the sequencing phase and the information regarding human genomic sequences has explosively grown,[1] the sequencing of human cDNAs has been increasing its importance because it plays an indispensable role in complementation of the information encoded in human genomic sequences. Taking this into consideration, we initiated a human cDNA sequencing project 4 years ago,[2] and have already determined more than 800 human cDNAs to date.[3] The notable point of our cDNA project is that we focus our sequencing efforts on the analysis of large cDNAs (> 4 kb). Recently, we have selected cDNA clones to be sequenced on the basis of their protein coding potentials, and cDNA clones which code for large proteins (> 50 kDa) in brain are the current targets in our cDNA project.[4]

As an extension of the preceding reports, we herein present the entire sequences of 100 new cDNA clones from brain cDNA libraries which encode large proteins *in vitro*. The specific features of the newly predicted protein sequences identified by the homology/motif analysis, the expression profiles, and the chromosomal locations of these 100 new cDNAs are also described. In particular, for 54 genes, the expression patterns in various regions of the central nervous system and fetal brain, besides 10 human tissues, were also examined. The expression profiling in these additional specimens was found to offer important pieces of information as a clue to identify their biological functions from a neuroscience viewpoint.

2. Materials and Methods

2.1. Source and screening of cDNA clones

cDNA clones were randomly isolated from size-fractionated human brain cDNA libraries Nos. 2 to 5 (average insert size = 3.9, 4.5, 5.3 and 6.1 kb, respectively).[4] As the first screening, an *in vitro* transcription/translation system was applied to select cDNA clones which have the coding potentials of proteins with apparent molecular mass larger than 50 kDa. In the next step, the clones with unidentified sequences at both

Communicated by Michio Oishi

* To whom correspondence should be addressed. Tel. +81-438-52-3913; Fax. +81-438-52-3914; E-mail: ohara@kazusa.or.jp

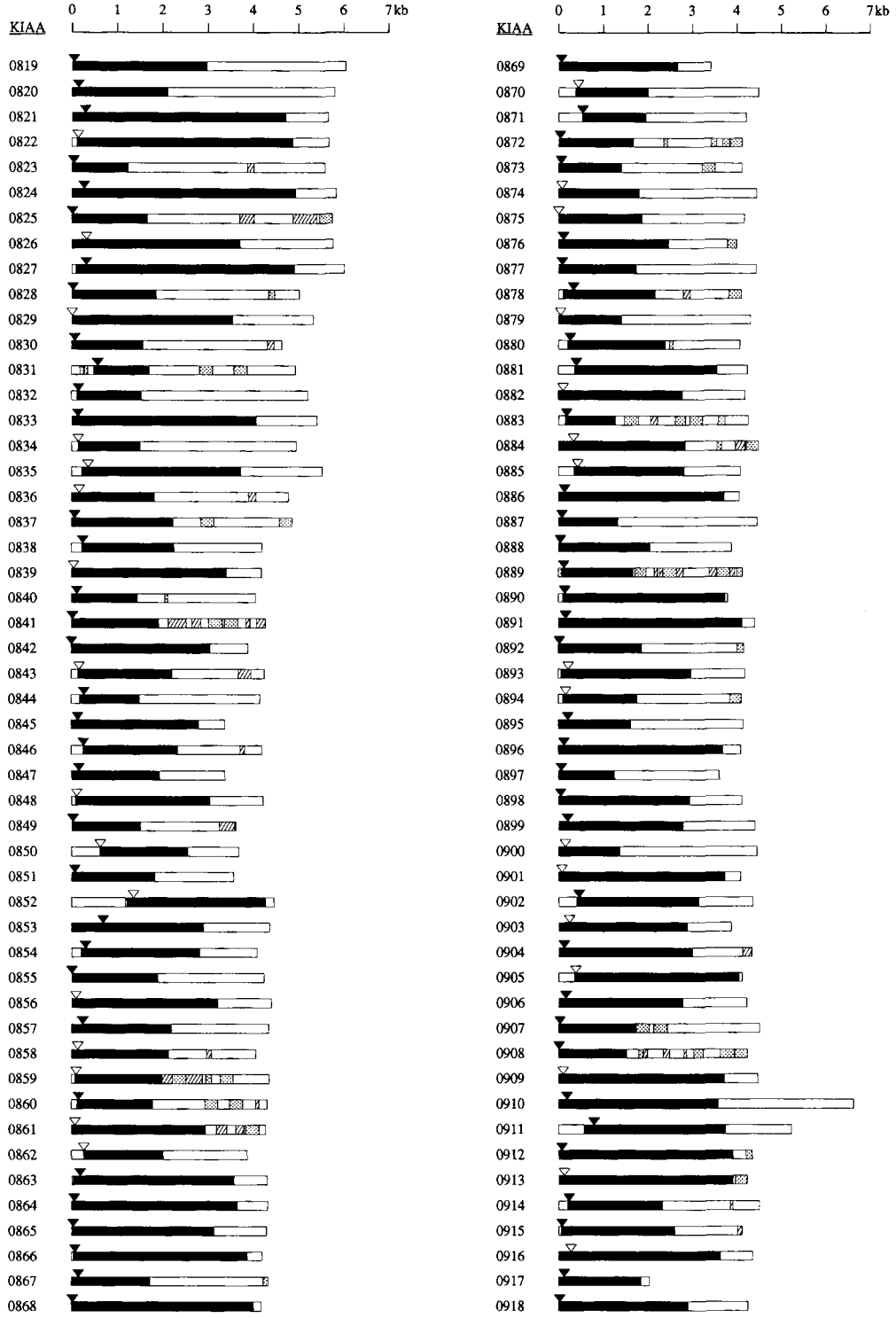


Figure 1. Physical maps of cDNA clones analyzed. The physical maps shown here were constructed on the basis of the sequence data of respective cDNA clones. The horizontal scale represents the cDNA length in kb, and the gene numbers corresponding to respective cDNAs are given on the left. The ORFs and untranslated regions are shown by solid and open boxes, respectively. The positions of the first ATG codons with or without the contexts of the Kozak's rule are indicated by solid and open triangles, respectively. RepeatMasker, which is a program that screens DNA sequences for interspersed repeats known to exist in mammalian genomes, was applied to detect repeat sequences in respective cDNA sequences (Smit, A.F.A. and Green, P., RepeatMasker at http://ftp.genome.washington.edu/RM/ RepeatMasker.html). Short interspersed nucleotide elements (SINEs) including Alu and MIRs sequences and other repetitive sequences thus detected are displayed by dotted and hatched boxes, respectively.

Table 1. Information of sequence data and chromosomal locations of the identified genes.

Gene number (KIAA)	Accession number[a]	cDNA length (bp)[b]	ORF length (amino acid residues)	Chromosomal location[c]
0819	AB020626	6,049	989	22
0820	AB020627	5,804	702	1*
0821	AB020628	5,659	1,474	19
0822	AB020629	5,677	1,581	17
0823	AB020630	5,597	412	20*
0824	AB020631	5,834	1,644	11
0825	AB020632	5,751	553	5
0826	AB020633	5,770	1,236	4*
0827	AB020634	6,019	1,531	16
0828	AB020635	5,025	619	7*
0829	AB020636	5,333	1,183	12*
0830	AB020637	4,638	523	11*
0831	AB020638	4,943	379	14
0832	AB020639	5,216	458	1*
0833	AB020640	5,414	1,353	1
0834	AB020641	4,957	451	7*
0835	AB020642	5,533	1,121	20
0836	AB020643	4,791	609	15*
0837	AB020644	4,868	745	5*
0838	AB020645	4,198	669	2
0839	AB020646	4,193	1,137	1*
0840	AB020647	4,059	483	5*
0841	AB020648	4,283	641	19
0842	AB020649	3,896	1,020	1
0843	AB020650	4,256	683	5
0844	AB020651	4,158	407	10*
0845	AB020652	3,385	933	22*
0846	AB020653	4,204	689	2*
0847	AB020654	3,386	645	15
0848	AB020655	4,220	977	3
0849	AB020656	3,622	504	16
0850	AB020657	3,682	642	1
0851	AB020658	3,572	587	3*
0852	AB020659	4,467	970	22[e]
0853	AB020660	4,363	967	13
0854	AB020661	4,089	837	8*
0855	AB020662	4,241	632	15
0856	AB020663	4,404	1,070	15*
0857	AB020664	4,341	733	2
0858	AB020665	4,059	711	13*
0859	AB020666	4,356	623	1*
0860	AB020667	4,313	541	20*
0861	AB020668	4,282	982	3
0862	AB020669	3,872	582	10*
0863	AB020670	4,313	1,131	18
0864	AB020671	4,319	1,218	17
0865	AB020672	4,300	1,045	13
0866	AB020673	4,199	1,266	16*
0867	AB020674	4,339	526	12
0868	AB020675	4,185	1,339	7*
0869	AB020676	3,408	888	5*
0870	AB020677	4,484	520	8
0871	AB020678	4,207	469	4*
0872	AB020679	4,118	547	16*
0873	AB020680	4,119	466	14*
0874	AB020681	4,440	601	18
0875	AB020682	4,168	621	12
0876	AB020683	3,997	819	19
0877	AB020684	4,436	580	7*
0878	AB020685	4,099	611	5*
0879	AB020686	4,312	453	6
0880	AB020687	4,068	709	11
0881	AB020688	4,242	1,049	16
0882	AB020689	4,186	924	4
0883	AB020690	4,253	364	8
0884	AB020691	4,487	943	14
0885	AB020692	4,076	798	1*
0886	AB020693	4,053	1,192	2
0887	AB020694	4,456	443	5
0888	AB020695	3,882	684	5*
0889	AB020696	4,122	518	20
0890	AB020697	3,800	1,194	3
0891	AB020698	4,401	1,371	3
0892	AB020699	4,164	621	19*
0893	AB020700	4,195	919	1*
0894	AB020701	4,113	533	10
0895	AB020702	4,155	540	7*
0896	AB020703	4,091	1,230	8*
0897	AB020704	3,605	419	1
0898	AB020705	4,111	979	17*
0899	AB020706	4,405	929	11
0900	AB020707	4,450	455	13*
0901	AB020708	4,078	1,215	X*
0902	AB020709	4,349	893	1
0903	AB020710	3,875	962	2
0904	AB020711	4,336	997	17
0905	AB020712	4,110	1,220	4
0906	AB020713	4,217	923	3*
0907	AB020714	4,500	564	1*
0908	AB020715	4,226	506	2
0909	AB020716	4,465	1,234	17
0910[d]	AB020717	6,624	1,193	21
0911[d]	AB020718	5,219	981	1*
0912[d]	AB020719	4,345	1,209	15
0913[d]	AB020720	4,227	819	10
0914[d]	AB020721	4,449	683	4
0915[d]	AB020722	4,065	777	17
0916[d]	AB020723	4,354	1,210	13*
0917[d]	AB020724	2,038	616	14*
0918[d]	AB020725	4,250	966	13

a) Accession numbers of DDBJ, EMBL and GenBank databases.
b) Values excluding poly(A) sequences.
c) Chromosome numbers identified by using GeneBridge 4 radiation hydrid panel unless specified. The chromosomal locations highlighted by asterisks were fetched from the UniGene database.
d) cDNA and ORF lengths were revised by direct analysis of the RT-PCR products.
e) Chromosome number determined by using CCR human-rodent hybrid panel.

Table 2. Functional classifications of the gene products based on homologies to known proteins and sequence motifs.

Functional category[a]	Gene number (KIAA)	Similarity class[b]	Homologous entry in the database[c]	Accession no.[d]	Identities (%)[e]	Overlap (amino acid residues)[e]
Cell signaling/communication	0821	H	calcium-independent alpha-latrotoxin receptor (R)	U72487	98.1	1474
	0823	W	MYPT2 (H)	AB003062	29.4	306
	0834	H	Pftaire-1 (M)	AF033655	97.9	429
	0846	R	ras guanyl releasing protein (R)	AF060819	51.5	718
	0856	R	X-like 1 protein (H)[f]	AJ005821	54.1	1084
	0857		none[g]			
	0861	R	KIAA0362 (H)	AB002360	43.0	881
	0862	I	Ras-binding protein SUR-8 (H)	U61957	99.8	582
	0868	R	contactin associated protein Caspr (H)	U87223	44.2	1322
	0871	W	RaP2 interacting protein 8 (H)	U93871	32.4	182
	0878	W	Caenorhabditis elegans cosmid C07D10 (Ce)[f]	U13072	24.6	281
	0880	R	prostaglandin transporter (R)	Q00910	42.8	678
	0881	R	S/T-protein kinase SULU (Ce)	P46549	37.4	962
	0882	W	MIC1 protein (Sc)	P53258	37.3	378
	0886	R	neuroendocrine-specific protein A (H)	A46583	33.4	685
	0893	R	Caenorhabditis elegans cosmid K06A5 (Ce)[f]	AF039038	39.3	379
	0894	W	SH3-containing protein p4015 (R)	AF026505	32.4	352
	0897	R	LAR-interacting protein LIP1a (H)	S55552	69.0	400
	0902	R	KIAA0403 (H)[f]	AB007863	38.1	270
	0904	R	protein kinase cdc2-like (H)	A38197	82.7	394
	0910	I	synaptojanin (H)	AF009039	97.3	1225
	0915	R	guanine nucleotide regulatory protein tim1 (H)	U02082	35.9	499
	0917	H	Sly1 protein (R)	JC4674	96.8	599
Cell structure/motility	0819	W	cell surface glycoprotein 1 precursor (Ct)	Q06852	24.6	691
	0820	H	dynamin 3 (R)	Q08877	96.2	679
	0843	R	actin-binding double-zinc-finger protein abLIM (H)	AF005654	49.8	652
	0845	I	neurofilament triplet H protein (H)	P12036	99.2	933
	0851	R	recessive suppressor of secretory defect (Sc)	P32368	35.4	591
	0853	W	trichohyalin (H)	P09406	25.9	247
	0855	W	cis-golgi matrix protein GM130 (R)	Q62839	29.4	884
	0865	W	ALR (H)	AF010403	28.8	417
	0866	H	myosin heavy chain, smooth muscle isoform (Rb)	P35748	97.6	1264
	0899	H	alpha-adaptin (R)	P18484	97.3	929
	0903	W	spectrin beta chain, erythrocyte (M)	P15508	21.2	754
	0905	W	Web1 protein (Sc)	P38968	24.2	1338
	0906	R	integral membrane glycoprotein GP210 precursor (R)	P11654	80.6	923
	0912	W	NuMA protein (X)	Y07624	22.7	948
Nucleic acid management	0824	I	PCF11p homolog (H)	AF046935	99.9	713
	0827	W	transcription factor NFATx (M)	D85612	28.6	961
	0829	H	TIP120 (R)	D87671	99.7	1183
	0832	R	steroid hormone receptor ERR2 (H)	P11475	77.0	435
	0835	I	myelin transcription factor 1 (H)	Q01538	93.5	536
	0854	R	homeotic protein zhx-1 (M)	JC4863	40.5	893
	0864	W	p116Rip (M)[f]	U73200	86.3	350
	0867	W	transcription factor-like protein 4 beta (M)	JC5332	26.9	227
	0885	H	UNR protein (R)[f]	P18395	98.6	798
	0890	W	pre-mRNA splicing factor RNA helicase PRP16 (Sc)	P15938	34.0	532
	0901	R	Caenorhabditis elegans cosmid F41H10 (Ce)[f]	U61954	38.7	791
	0916	I	protein associated with Myc (H)	AF075587	99.9	1210
Metabolism	0828	I	S-adenosyl homocysteine hydrolase homolog (H)	U82761	91.8	500
	0836	H	C5-glucuronyl epimerase (B)	AF003927	98.9	444
	0837	H	long-chain-fatty-acid--CoA ligase, brain isozyme (R)	P33124	90.7	697
	0838	H	glutaminase, kidney isoform precursor (R)	P13264	93.9	674
	0879	R	plasma-cell membrane glycoprotein PC-1 (M)	P06802	39.5	390
Protein management	0887	W	probable membrane protein YDL091c (Sc)[f]	S67627	28.5	302
	0891	R	ubiquitin protease Unph (H)	U20657	35.2	824
	0896	I	100 kD protein (H)	Q62671	97.8	889
Unclassified	0822	W	Caenorhabditis elegans cosmid F12B6 (Ce)	AF003138	24.6	1416
	0826	W	Caenorhabditis elegans cosmid F21H11 (Ce)	U11279	21.9	356
	0830	R	K123 protein (C)	Y14970	32.3	263
	0833	W	Caenorhabditis elegans cosmid T05C1 (Ce)	U28992	26.5	181
	0839	W	Caenorhabditis elegans cosmid T22C1 (Ce)	Z75550	22.8	334
	0840	W	hypothetical 54.9 kD protein C02F5.7 (Ce)	P34284	29.7	384
	0849	R	F40F12.5 protein (Ce)	S42834	33.0	455
	0850	W	Caenorhabditis elegans cosmid R09A8 (Ce)	Z68009	28.9	724
	0852	W	KIAA0136 (H)	Q14149	24.9	1013
	0858	I	zinc-finger domain-containing protein (H)	U90654	100.0	341
	0859	R	Caenorhabditis elegans cosmid C01B10 (Ce)	U58757	35.5	617
	0863	R	hypocotil specific gene (Dc)	AB000505	33.2	349
	0872	W	Caenorhabditis elegans cosmid F41H10 (Ce)	U61954	28.4	443
	0875	W	hypothetical protein HI 1558 (Hi)	P45252	25.9	185
	0876	W	peregrin (H)	P55201	28.0	432
	0877	R	DPY-19 protein (Ce)	P34413	41.1	555
	0884	H	tulip 1 (R)	AF041106	93.5	230
	0888	W	probable chitin biosynthesis protein C6G9.12 (Sp)	Q92357	24.8	322
	0892	W	Caenorhabditis elegans cosmid C09H6 (Ce)	Z81466	23.0	600
	0898	W	Ro protein (M)	L27990	25.1	263
	0900	R	KIAA0269 (H)	Q92558	47.3	514
	0907	W	hypothetical 59.0 kD protein (Sp)	Q09911	25.8	461
	0909	W	Caenorhabditis elegans cosmid T05C1 (Ce)	U28992	33.2	223

	0911	W	Caenorhabditis elegans cosmid B0034 (Ce)	U23528	29.4	991
	0913	W	Caenorhabditis elegans cosmid R09E10 (Ce)	Z70287	34.8	181
	0914	R	clone 23909 (H)	U79304	40.2	261
No homology	0825		none			
	0831		none			
	0841		none			
	0842		none			
	0844		none			
	0847		none			
	0848		none			
	0860		none			
	0869		none			
	0870		none			
	0873		none			
	0874		none			
	0883		none			
	0889		none			
	0895		none			
	0908		none			
	0918		none			

a) Classifications based on the annotations of their homologous protein entries in the databases.
b) The gene products were grouped into four similarity classes according to the sequence identities obtained by the GAP program: I, identical to known human gene products (sequence identity, > 90%); H, homologous to known non-human gene products (sequence identity, > 90%); R, related to some known gene products (sequence identity, 30 to 90%); W, very weakly related to known gene products (sequence identity, < 30%).
c) Organisms in which these entries were identified are given in parentheses: B, bovine; C, chichen; Ce, *Caenorhabditis elegans*; Ct, *Clostridium thermocellum*; Dc, *Daucus carota*; H, human; Hi, *Haemophilus influenzae*; M, mouse; R, rat; Rb, Rabbit; Sp, *Schizosaccharomyces pombe*; X, *Xenopus laevis*.
d) Accession numbers of homologous entries in DDBJ/EMBL/GenBank/OWL/SWISS-PROT/PIR database are shown.
e) The values were obtained by the FASTA program.
f) Classifications based on the sequence motifs or domains.

ends were chosen by single-pass sequencing and homology search against GenBank database (release 102.0) excluding expressed sequence tags and genomic sequences.[3] Although cDNAs with unidentified sequences were analyzed as a general rule, cDNA clones with much larger open reading frames (ORFs) than the registered ones in the public databases were also included in this report as exceptions.

2.2. *Gene expression profiles*

Expression profiles of the newly identified genes were examined by reverse transcription-coupled polymerase chain reaction (RT-PCR), products of which were quantified by using enzyme-linked immunosorbent assay (ELISA) as described previously.[3] On the basis of ELISA control curves using PCR products derived from serial dilutions of a known amount of the authentic plasmids, the ELISA data were converted to the mRNA levels expressed as equivalent amounts of the cDNA plasmid. The digitized mRNA levels calculated using a software package, SOFTmax PRO (Molecular Device, Co.), were then displayed by color codes for facilitating survey of many gene expression profiles at a glance. In this study, poly(A)$^+$ RNAs from 8 brain regions (amygdala, corpus callosum, cerebellum, caudate nucleus, hippocampus, substantia nigra, subthalamic nucleus, and thalamus), spinal cord, fetal brain, and fetal liver were subjected to the RT-PCR ELISA in addition to those from 10 adult human tissues as described previously.[3] All the poly(A)$^+$ RNAs were purchased from CLONTECH Laboratories, Inc. (Palo Alto, CA, USA).

2.3. *Other methods*

DNA sequencing and homology searches of the predicted protein-coding sequences were carried out as described previously.[3,4] When the possibility of spurious interruption of ORF was noticed, the interruption was experimentally revised by direct sequencing of the RT-PCR products.[5] Chromosomal locations of newly identified genes were determined by using human-rodent hybrid panels, GeneBridge 4 (Research Genetics Inc., USA) or CCR (Coriell Cell Repositories, USA), if their mapping data were not available in the UniGene database (http://www.ncbi.nlm.nih.gov/UniGene).[3] For genes whose chromosomal locations were described in the UniGene database, we did not perform the radiation hybrid mapping experiments. In this case, we confirmed that the primer sets used in the determination of chromosomal location in the UniGene database were consistent with the sequences of the genes we determined. The actual primer sequences and the PCR conditions used for the radiation hybrid mapping are accessible through the World Wide Web at http://www.kazusa.or.jp.

Table 3. The newly identified genes in paralogous relationship with genes characterized by our cDNA project.

New gene	Paralogous gene	Accession no.[a)]	Identities (%)[b)]	Corresponding gene product
KIAA0821	KIAA0768	AB018311	53.0	latrophilin-related protein 1[11]
	KIAA0786	AB018329	64.7	latrophilin-related protein 1
KIAA0829	KIAA0667	AB014567	57.5	TBP-interacting protein TIP120[12]
KIAA0833	KIAA0908[c)]	AB020715	41.7	uncharacterized
KIAA0835	KIAA0535	AB011107	51.3	myelin transcription factor[13]
KIAA0843	KIAA0059	D31883	56.5	actin-binding protein[14]
KIAA0850	KIAA0132	D50922	30.9	ring canal protein[15]
KIAA0854	KIAA0395	AB007855	38.0	homeotic protein zhx-1[16]
KIAA0861	KIAA0362	AB002360	42.7	guanine nucleotide exchange factor Dbs[17]
KIAA0866	KIAA0216	D86970	33.1	smooth muscle myosin heavy chain[18]
	KIAA0389	AB002387	31.1	smooth muscle myosin heavy chain
KIAA0875	KIAA0677	AB014577	43.2	uncharacterized
KIAA0881	KIAA0676	AB014576	61.1	uncharacterized
KIAA0890	KIAA0529	AB011101	32.0	uncharacterized
KIAA0896	KIAA0654	AB014554	65.8	LAR-interacting protein LIP1a[19]
KIAA0899	KIAA0269	D87459	51.0	uncharacterized
KIAA0901	KIAA0403	AB007863	34.6	uncharacterized
KIAA0909	KIAA0348	AB002346	45.7	synaptojanin[20]
KIAA0918	KIAA0848[c)]	AB020655	43.7	uncharacterized

a) Accession numbers of paralogous genes in DDBJ/EMBL/GenBank database are shown.
b) The values of the overall identities of amino acid residues were obtained by the GAP program.
c) These genes are reported in this paper.

3. Results and Discussion

3.1. Sequence analysis and prediction of protein-coding regions in cDNA clones

cDNA clones were selected from a set of size-fractionated cDNA libraries under the criteria described in the previous studies;[3] they are uncharacterized in the public databases and can direct synthesis of proteins larger than 50 kDa *in vitro*. One hundred clones thus selected were subjected to sequencing of entire inserts. Some clones seemed to carry spurious coding interruption caused by errors of the reverse transcriptase or retained intron sequences. For these cases, the regions causing ORF interruption were examined by direct sequencing of the RT-PCR products. It should be noted that only the main RT-PCR products were taken into consideration. According to the results of these confirmations, the spurious interruptions were found in the following clones: the ORF in KIAA0915 was interrupted by a 49-bp deletion; KIAA0916 carried a nonsense mutation in the ORF; KIAA0917 was found to carry relatively long insertions probably corresponding to intronic sequences; the ORFs in 5 clones (KIAA0910-0913 and KIAA0918) were frame-shifted by a short insertion or deletion (< 5 nucleotide residues). For those genes, the revised sequences by the RT-PCR experiments, not the actual cloned cDNA sequences, were deposited to the GenBank/EMBL/DDBJ databases and used for prediction of protein coding sequences. In particular, KIAA0914 was found to generate two alternative forms without causing the ORF interruption, and the nucleotide sequence with a 42-bp insertion was deposited to the public databases. The sequence data revealed that the average sizes of these cDNA inserts and of their ORFs were 4.4 kb and 2.5 kb (corresponding to 831 amino acid residues), respectively. Physical maps of the 100 cDNA clones analyzed are shown in Fig. 1, where the ORFs and the first ATG codons in respective ORFs are indicated by solid boxes and triangles, respectively. The in-frame termination codons upstream of the first ATG codon were identified in 44 clones, among which 27 clones carried the ATG codon within the context of Kozak's rule.[6] In Fig. 1, short interspersed nucleotide elements (Alu and MIRs sequences) and other repetitive sequences detected using the RepeatMasker program are indicated by dotted and hatched boxes, respectively. Table 1 lists the gene codes (KIAA numbers), the accession numbers of the nucleotide sequences in GenBank/EMBL/DDBJ databases, the sizes of the cDNA inserts and the identified ORF, and the chromosomal locations of the respective genes. The chromosomal locations of 44 genes, which are highlighted by asterisks, were fetched from the UniGene database while the remaining 56 chromosomal locations were experimentally determined in this study.

3.2. Functional classification of predicted gene products

To classify the gene products predicted from the cDNA sequences according to their possible functions,

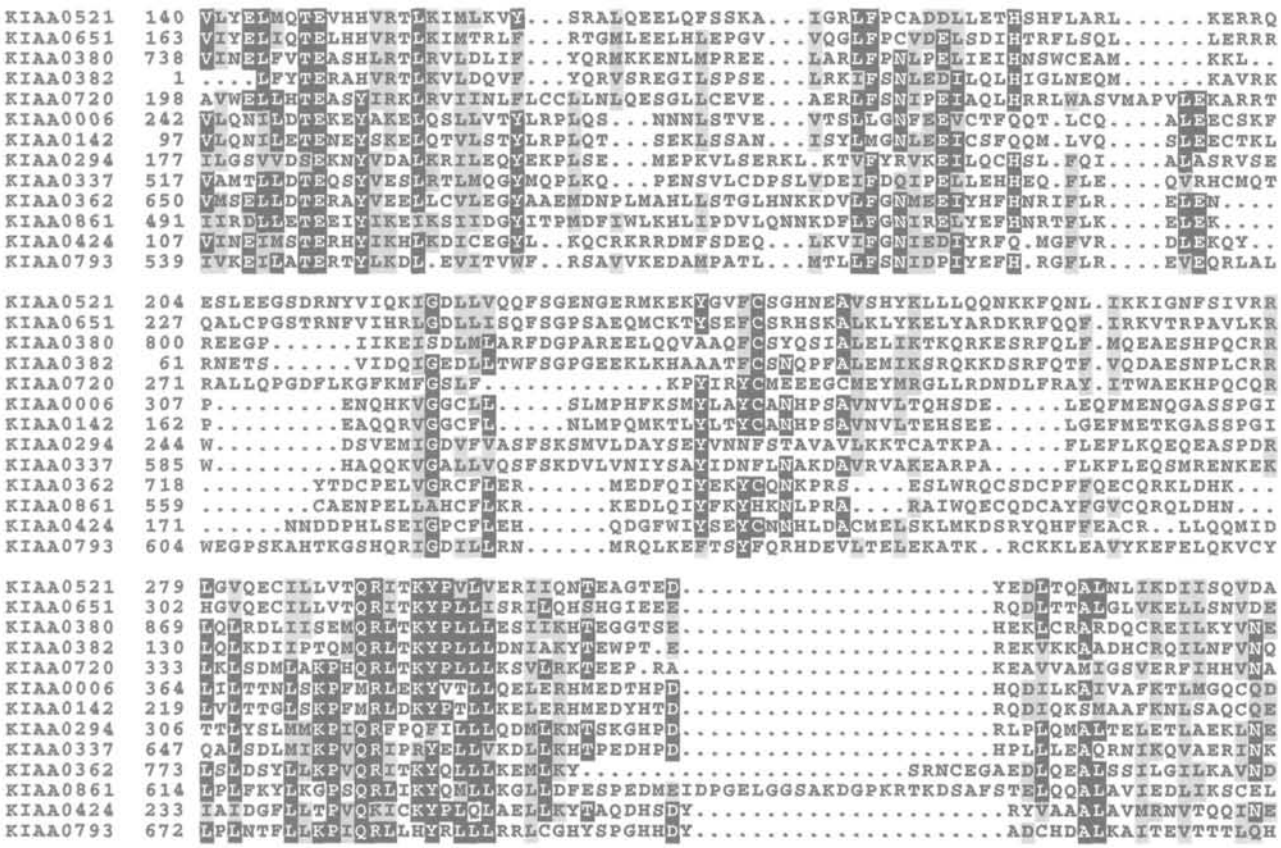


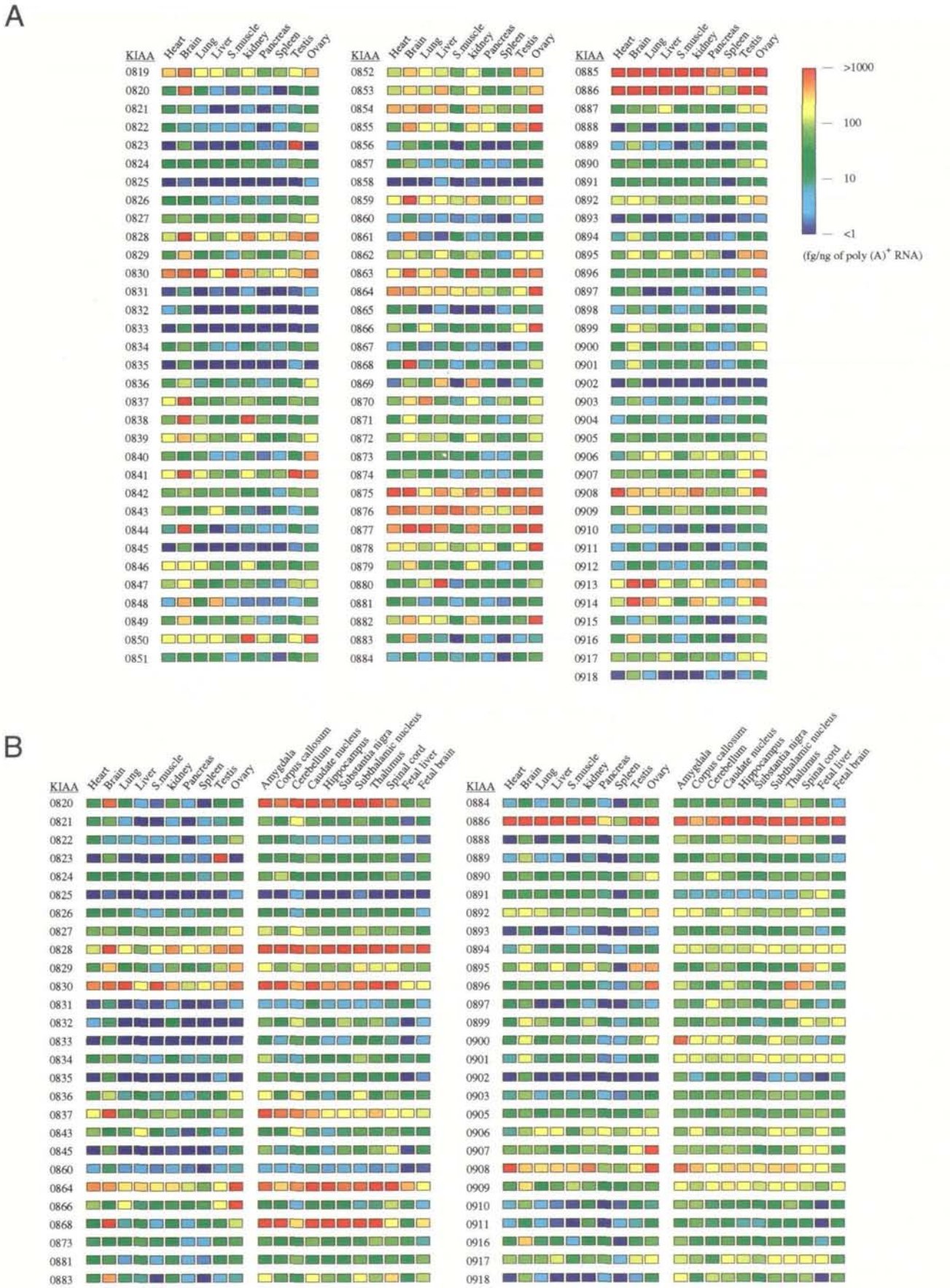
Figure 2. Multiple sequence alignment of the 13 DH domains. The DH domains of 13 KIAA genes were aligned by the PILEUP program. Identical residues are illustrated in black while conserved residues are in gray. Numerals represent the number of amino acid residues in respective genes. Gene names are given at the left.

the similarities of the sequences were examined against DNA and protein databases [GenBank (release 109.0), OWL (release 30.3) databases] by using the FASTA program of Wisconsin Sequence Analysis Package™ (version 8; Genetics Computer Group, Inc., USA). Motifs and profile entries in the PROSITE database (release 15.0) were also searched for by using the MOTIFS program in the Wisconsin Sequence Analysis Package and pftools program (ftp://ulrec3.unil.ch/pub/pftools), respectively. Furthermore, the Pfam database (release 3.2; http://pfam.wustl.edu), which is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains,[7] were also used for identification of functional domains in the predicted protein sequences (http://hmmer.wustl.edu). Homology searches against these databases revealed that the predicted coding sequences of 83 genes were found to exhibit significant similarities to those of known genes or specific domains/motifs, and 59% of them were classified into protein groups functionally related to cell signaling/communication, cell structure/motility and nucleic acid management. The functional classifications of these newly identified genes on the basis of this homology/motif analysis are summarized in Table 2.

Interesting features to be noted are summarized below.

1. Nineteen newly identified genes constitute 17 independent paralogous groups together with the genes characterized in our cDNA project (Table 3). For this analysis, genes which exhibit significant similarities throughout the protein-coding sequences, not in distinct domains or motifs, have been assigned as those with a paralogous relationship. Among these paralogous groups, genes in 7 groups were annotated "uncharacterized."

2. The membrane-spanning regions in the gene products predicted from the cDNA sequences were predicted by SOSUI, which is a program for classification and secondary structure prediction for membrane proteins.[8] By using this program, more than 8 membrane-spanning segments in single polypeptides were predicted in four gene products analyzed in this paper (KIAA0821, 8 segments; KIAA0822, 14 segments; KIAA0877, 10 segments; KIAA0880, 8 segments). These genes are likely to encode receptors or ion channels or transporters. In fact, KIAA0821 and KIAA0880 showed significant sequence similarities against calcium-independent alpha-latrotoxin



receptor and prostaglandin transporter, respectively (Table 2).

3. A protein domain search against the Pfam database[7] revealed that the Dbl homology (DH) domain[9] (HMM file, PF00621.pfam [RhoGEF]) and *Hect* domain[10] (HMM file, PF00632.pfam) are present in KIAA0861 and KIAA0896, respectively. Both domains were frequently found in the predicted gene products in our cDNA project. To examine the relative similarity among the DH-domain containing proteins, 13 DH domains in KIAA gene products were aligned by the PILEUP program in the Wisconsin Sequence Analysis Package (Fig. 2). As shown in Fig. 2, the region with the highest degree of similarity within the DH domain was characterized by a consensus sequence, Gln-Arg-Leu/Ile-Xaa-Lys-Tyr-Pro-Leu-Leu-Leu. Although the sequence identity of the DH domains in these proteins was not extremely high, the DH domains in all these predicted proteins except KIAA0337 are followed by Pleckstrin-homologous (PH) domain, which is thought to be involved in protein-protein interaction. The sequence diversity of the DH-containing human proteins suggested the existence of multiple signaling cascades in higher eukaryote since the DH domains are found in only four predicted proteins (ROM1, YLR371W, CDC24 and FUS2) in *Saccharomyces cerevisiae*.

4. Protein motif search against PROSITE database revealed that protein kinase signatures are present in three genes (KIAA0834, KIAA881 and KIAA0904).

3.3. Expression profiles of predicted genes

Figure 3A shows the tissue expression profiles of 100 human cDNAs newly identified in this study. Although RT-PCR ELISA cannot be highly accurate in quantification in an analytical sense, we consider it reliable enough as a screening method of genes according to their expression patterns.[3] However, from a neuroscientific viewpoint, the expression profiles among 10 tissues are not always satisfactory for extracting clues to identify their biological significance. Thus, we examined whether or not more neuroscientifically relevant information can be obtained by expanding the range of samples subjected to expression profiling. In this context, we analyzed the expression patterns of 54 randomly selected new genes in 8 brain regions (amygdala, corpus callosum, cerebellum, caudate nucleus, hippocampus, substantia nigra, subthalamic nucleus, and thalamus), spinal cord, and fetal brain in addition to 10 human tissues. In Fig. 3B, the expression profiles of these 54 genes in various regions of the central nervous system and fetal brain are shown together with those in 10 human tissues. Although mRNAs in whole brain also contains those from these various brain regions, the majority of whole brain mRNAs is derived from cerebrum and thus the mRNA levels in whole brain is expected to be similar to those in the cerebrum. Thus, even if the mRNA level of a gene of interest in adult whole brain is low, it can happen that a particular region of the central nervous system or fetal brain contained a significant amount of the mRNA. Figure 3B shows that several genes (e.g., KIAA0821, KIAA0832, KIAA0836 and so on) were expressed in this way. Such brain region-specific or developmentally controlled genes are as interesting as brain-specific genes in a biological sense. While these genes expressed in a regionally or developmentally specific way could not be identified in the previous expression profiling, the inclusion of samples derived from various regions of the central nervous system and from fetal brain in the expression profiling would prevent us from overlooking these biologically interesting genes. Therefore, together with the specific sequence features of the predicted gene products and their chromosomal locations, the expression profiles of new genes in adult tissues, various regions of the central nervous system and fetal brain would provide important clues as to which genes are to be further investigated.

Acknowledgments: This project was supported by grants from the Kazusa DNA Research Institute. We thank Tomomi Tajino, Keishi Ozawa, Tomomi Kato, Kazuhiro Sato, Akiko Ukigai, Emiko Suzuki, Kazuko Yamada, Kiyoe Sumi, Takashi Watanabe, Naoko Suzuki, Kozue Kaneko, Naoko Shibano, Taneaki Tsugane and Hisako Takahashi for their technical assistance.

References

1. Rowen, L., Mahairas, G., and Hood, L. 1997, Sequencing the human genome, *Science*, **278**, 605–607.
2. Nomura, N., Miyajima, N., Sazuka, T. et al. 1994, Prediction of the coding sequences of unidentified human genes. I. The coding sequences of 40 new genes (KIAA0001-0040) deduced by analysis of randomly sampled cDNA clones from human immature myeloid cell line KG-1,

Figure 3. Expression profiles of 100 newly identified genes examined by RT-PCR ELISA. The tissue expression levels of the 100 human genes were analyzed by using the RT-PCR ELISA. Panel A: Expression profiles among 10 human adult tissues. Gene names are given as KIAA numbers at the left side of each set of color codes. Tissue names are indicated on above the top sets of color codes. A color conversion panel shown in the right side was used for displaying mRNA levels as color codes. The mRNA levels are expressed in equivalent amounts (fg) of the authentic cDNA plasmids in 1 ng of starting poly(A)$^+$ RNAs. Panel B: Brain regional and developmental expression profiles of randomly selected 54 genes analyzed above. Besides 10 tissues, 9 regions of the adult central nervous system (amygdala, corpus callosum, cerebellum, caudate nucleus, hippocampus, substantia nigra, subthalamic nucleus, thalamus, and spinal cord) and fetal brain were included in the expression profiling. As a control, mRNA levels in fetal liver were also examined. Gene names and sample names of mRNA origins are indicated in the same way as in panel A.

DNA Res., **1**, 27–35.
3. Nagase, T., Ishikawa, K.-I., Suyama, M. et al. 1998, Prediction of the coding sequences of unidentified human genes. XI. The complete sequences of 100 new cDNA clones from brain which code for large proteins *in vitro*, *DNA Res.*, **5**, 277–276.
4. Ohara, O., Nagase, T., Ishikawa, K.-I. et al. 1997, Construction and characterization of human brain cDNA libraries suitable for analysis of cDNA clones encoding relatively large proteins, *DNA Res.*, **4**, 53–59.
5. Ishikawa, K.-I., Nagase, T., Nakajima, D. et al. 1997, Prediction of the coding sequences of unidentified human genes. VIII. 78 new cDNA clones from brain which code for large proteins *in vitro*, *DNA Res.*, **4**, 307–313.
6. Kozak, M. 1996, Interpreting cDNA sequences: some insights from studies on translation, *Mammalian Genome*, **7**, 563–574.
7. Sonnhammer, E. L., Eddy, S. R., Birney, E., Bateman, A., and Durbin, R. 1998, Pfam: multiple sequence alignments and HMM-profiles of protein domains, *Nucleic Acids Res.*, **26**, 320–322.
8. Hirokawa, T., Boon-Chieng, S., and Mitaku, S. 1998, SOSUI: classification and secondary structure prediction system for membrane proteins, *Bioinformatics*, **14**, 378–379.
9. Chan, A. M.-L., McGovern, E. S., Catalano, G., Fleming, T. P., and Miki, T. 1994, Expression cDNA cloning of a novel oncogene with sequence similarity toregulators of small GTP-binding proteins, *Oncogene*, **9**, 1057–1063.
10. Huibregtse, J. M., Scheffner, M., Beaudenon, S., and Howley, P. M. 1995, A family of proteins structurally and functionally related to the E6-AP ubiquitin-protein ligase, *Proc. Natl. Acad. Sci. USA*, **92**, 2563–2567.
11. Lelianova, V. G., Davletov, B. A., Sterling, A. et al. 1997, α-latrotoxin receptor, latrophilin, is a novel member of the secretin family of G protein-coupled receptors, *J. Biol. Chem.*, **272**, 21504–21508.
12. Yogosawa, S., Makino, Y., Yoshida, T., Kishimoto, T., Muramatsu, M., and Tamura, T. 1996, Molecular cloning of a novel 120-kDa TBP-interacting protein, *Biochem. Biophys. Res. Commun.*, **229**, 612–617.
13. Kim, J. G., Armstrong, R. C., v Agoston, D. et al. 1997, Myelin transcription factor 1 (Myt1) of the oligodendrocyte lineage, along with a closely related CCHC zinc finger, is expressed in developing neurons in the mammalian central nervous system, *J. Neurosci. Res.*, **50**, 272–290.
14. Roof, D. J., Hayes, A., Adamian, M., Chishti, A. H., and Li, T. 1997, Molecular characterization of abLIM, a novel actin-binding and double zinc finger protein, *J. Cell. Biol.*, **138**, 575–588.
15. Xue, F. and Cooley, L. 1993, kelch encodes a component of intercellular bridges in *Drosophila* egg chambers, *Cell*, **72**, 681–693.
16. Barthelemy, I., Carramolino, L., Gutierrez, J., Barbero, J. L., Marquez, G., and Zaballos, A. 1996, zhx-1: a novel mouse homeodomain protein containing two zinc-fingers and five homeodomains, *Biochem. Biophys. Res. Commun.*, **224**, 870–876.
17. Whitehead, I., Kirk, H., and Kay, R. 1995, Retroviral transduction and oncogenic selection of a cDNA encoding Dbs, a homolog of the Dbl guanine nucleotide exchange factor, *Oncogene*, **10**, 713–721.
18. Babij, P., Kelly, C., and Periasamy, M. 1991, Characterization of a mammalian smooth muscle myosin heavy-chain gene: complete nucleotide and protein coding sequence and analysis of the 5′ end of the gene, *Proc. Natl. Acad. Sci. USA*, **88**, 10676–10680.
19. Serra-Pages, C., Kedersha, N. L., Fazikas, L., Medley, Q., Debant, A., and Streuli, M. 1995, The LAR transmembrane protein tyrosine phosphatase and a coiled-coil LAR-interacting protein co-localize at focal adhesions, *EMBO J.*, **14**, 2827–2838.
20. Haffner, C., Takei, K., Chen, H. et al. 1997, Synaptojanin 1: localization on coated endocytic intermediates in nerve terminals and interaction of its 170 kDa isoform with Eps15, *FEBS Lett.*, **419**, 175–180.