DNA RESEARCH 4, 141–150 (1997)

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

Takahiro NAGASE, Ken-ichi ISHIKAWA, Daisuke NAKAJIMA, Miki OHIRA, Naohiko SEKI, Nobuyuki MIYAJIMA, Ayako TANAKA, Hirokazu KOTANI, Nobuo NOMURA, and Osamu OHARA*

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

(Received 21 March 1997)

Abstract

In this series of projects of sequencing human cDNA clones which correspond to relatively long transcripts, we newly determined the entire sequences of 100 cDNA clones which were screened on the basis of the potentiality of coding for large proteins *in vitro*. The cDNA libraries used were the fractions with average insert sizes from 5.3 to 7.0 kb of the size-fractionated cDNA libraries from human brain. The randomly sampled clones were single-pass sequenced from both the ends to select clones that are not registered in the public database. Then their protein-coding potentialities were examined by an *in vitro* transcription/translation system, and the clones that generated proteins larger than 60 kDa were entirely sequenced. Each clone gave a distinct open reading frame (ORF), and the length of the ORF was roughly coincident with the approximate molecular mass of the *in vitro* product estimated from its mobility on SDS-polyacrylamide gel electrophoresis. The average size of the cDNA clones sequenced was 6.1 kb, and that of the ORFs corresponded to 1200 amino acid residues. By computer-assisted analysis of the sequences with DNA and protein-motif databases (GenBank and PROSITE databases), the functions of at least 73% of the gene products could be anticipated, and 88% of them (the products of 64 clones) were assigned to the functional categories of proteins relating to cell signaling/communication, nucleic acid managing, and cell structure/motility. The expression profiles in a variety of tissues and chromosomal locations of the sequenced clones have been determined. According to the expression spectra, approximately 11 genes appeared to be predominantly expressed in brain. Most of the remaining genes were categorized into one of the following classes: either the expression occurs in a limited number of tissues (31 genes) or the expression occurs ubiquitously in all but a few tissues (47 genes).

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

1. Introduction

To accumulate information on the protein-coding sequences of unidentified human genes, we have implemented a project for sequencing the entire cDNA clones which are derived from relatively long transcripts.[1] We previously surveyed annotations of the entries of large proteins (> 1000 amino acid residues) in the public databases, and found that, even though the number of the entries is not large, more than 80% of large proteins in the databases are closely related to biologically important functions including some disorders in mammals.[2] Since the coding capacity of cDNA increases with its size, selective characterization of cDNA clones corresponding to long transcripts should be an effective approach to discover new genes of biological importance.

In the preceding paper, we constructed a set of size-fractionated cDNA libraries from human brain and demonstrated that an *in vitro* transcription/translation system could be used for selection of cDNA clones encoding large proteins.[2] We now report the coding sequences of 100 new cDNA clones selected on the basis of the protein-coding potentiality *in vitro*. cDNA clones were randomly sampled from the size-fractionated libraries from human brain, unidentified clones were selected on the basis of their terminal sequences, the clones encoding large proteins were further selected by analysis of the protein-coding potentialities in the *in vitro* system, and the clones that gave the *in vitro* products larger

Communicated by Mituru Takanami

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

than 60 kDa were entirely sequenced. Each clone gave a distinct open reading frame (ORF), and the length of the ORF was roughly coincident with the approximate molecular mass of the *in vitro* product estimated from its mobility on SDS-polyacrylamide gel electrophoresis (PAGE). The sequences of the predicted proteins were computer searched against the databases of protein primary structures and motifs, and the function of 73% of the gene products could be anticipated from annotations of known entries in the databases.

2. Materials and Methods

2.1. The source and screening of cDNA clones

cDNA clones were randomly selected from fractions 4 to 6 (average insert size = 5.3, 6.1, and 7.0 kb) of the size-fractionated cDNA libraries from human brain previously constructed.[2] All the clones were subjected to single-pass sequencing from both the ends using M13 forward and reverse primers. The terminal sequences were computer searched using the MPSRCH program[3] against a subset of the GenBank database (release 93.0) which contained primate DNA sequences but no expression sequence tags (ESTs), and those with scores of lower than 250 were categorized "unregistered." The coding potentiality of cDNA clones carrying unregistered sequences at both the ends were examined using *in vitro* transcription/translation analysis.[2] Prestained protein size markers (molecular masses from 202 kDa to 6.9 kDa; Kaleidoscope size markers from Bio-Rad Laboratories, USA) were used to estimate the apparent molecular masses of the *in vitro* products. Since multiple discrete bands are usually generated from a single clone (see Fig. 5 in ref. 2), the size of the largest band was estimated, and the clones gaving products larger than 60 kDa on SDS-PAGE were entirely sequenced by methods described previously.

2.2. Similarity analysis of the predicted protein-coding sequences

The entire cDNA sequences deduced were routinely analyzed with the GCG software package.[4] The similarity of the predicted coding sequences to known ones was examined using the FASTA program against the non-redundant protein database, OWL (release 29.1), and further checked by the GAP program. The motif search was performed by the MOTIFS program using the PROSITE database (release 13.0).

2.3. Expression profiles of the sequenced cDNA clones

The expression profiles of the cDNA clones sequenced in various tissues were monitored by reverse transcription followed by the polymerase chain reaction (RT-PCR). The cDNA templates for RT-PCR were synthesized from 1 μg of human poly $(A)^+$ RNA (CLONTECH Laboratories, Inc., USA) using excess amounts of Superscript II reverse transcriptase (Gibco BRL, USA) and random hexamer primers. After synthesis of the first-strand cDNA, the remaining RNA in the reaction mixture was degraded with RNase A, and the resulting cDNA templates were recovered by phenol extraction followed by ethanol precipitation. An aliquot of the cDNA template mixture [2 μl, corresponding to 1 ng of the starting poly $(A)^+$ RNA] was subjected to PCR using LA *Taq* DNA polymerase (0.5 units, Takara Shuzo Co., Ltd., Japan) in 10 μl with the DNA Thermal Cycler PJ9600 (Perkin Elmer, USA) and a set of primers specific to a gene of interest. As a positive control, a primer set for the glyceraldehyde-3-phosphate dehydrogenase (G3PDH) gene was used. Unless otherwise stated, the thermal cycling conditions used were as follows: The first denaturation at 95 °C for 1 min; 30 cycles of 0.5-min denaturation at 95 °C, 0.5-min primer annealing at 55 °C, and 1-min polymerization at 72 °C; and the last extension at 72 °C for 6 min. To check the efficiencies of PCR for individual genes, control reactions were conducted in which the PCR products were generated from serial dilutions of each cDNA clone (0.1 fg to 1 pg in 10 μl PCR mixture). The whole PCR products were loaded on 2.5% NuSieve GTG agarose gel (FMC BioProducts, USA) and detected by fluorescent staining with ethidium bromide. The gel images were recorded with a gel print 2000i/VGA (BioImage, USA). The PCR primer sets specific to individual genes were identical to those used for radiation hybrid mapping.

2.4. Chromosomal location

The cDNA clones entirely sequenced were mapped along chromosomes by using the GeneBridge 4 radiation hybrid panel (Research Genetics, Inc., USA). The details of the experimental conditions for the radiation hybrid mapping of respective clones are described elsewhere. The software for analysis of the results of the panel was obtained via ftp at ftp://genome.wi.mit.edu/distribution/software/rhmapper/.

3. Results and Discussion

3.1. Sequence analyses and prediction of the protein-coding regions of cDNA clones

Starting from the randomly sampled cDNA clones from the size-fractionated cDNA libraries of human brain, the clones to be sequenced were screened by the following two steps:

(1) Selection of unidentified cDNA clones by comparison of the single-pass sequences from both the ends with the primate DNA sequences except for ESTs registered in the database.

(2) selection of the clones encoding proteins larger than

60 kDa by analysis of the protein-coding potentiality *in vitro*.

Approximately 80% of the randomly sampled clones were sorted to the unregistered class, and approximately 20% of the non-redundant unregistered clones were found to have the potentiality of coding for proteins larger than 60 kDa. The entire sequences of the finally selected clones were deduced, and their sequence features were analyzed by computer. In addition, the chromosomal mapping of the sequenced clones using the radiation-hybrid panel and analysis of their expression profiles in 14 different tissues were performed. 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. Table 1 lists the lengths of inserts, the ORF lengths, the apparent molecular masses of the largest *in vitro* products for respective cDNA clones, and the chromosomal locations of the respective clones. The average size of the cDNA inserts was 6.1 kb and that of the predicted ORFs corresponded to 1200 amino acid residues. Although we listed the bands of the molecular masses larger than 100 kDa as > 100 kDa in Table 1 because of inaccuracy of the estimation, the values deduced from their relative mobilities on SDS-PAGE appeared to roughly coincide with the length of ORFs within a range of estimation errors. The result indicated that the selection of cDNA clones based on analysis of the *in vitro* products works well and eliminates those carrying short ORFs. The in-frame termination codons upstream of the first ATG codon were identified in 49 clones, in which 34 clones carried the ATG codon within the contexts of Kozak's rule: RNNATGG, RNNATGY, and YNNATGG, where R and Y indicate purine and pyrimidine residues, respectively, and ATG codons are underlined. In Fig. 1, such ATG codons surrounded by the contexts of Kozak's rule[5] are shown by solid triangles. In the remaining clones (51 clones), the in-frame termination codons were not identified in the 5′-untranslated region (5′-UTR), whereas distinct protein products were generated in the *in vitro* system. The presence of in-frame termination codons in the 5′-UTR is not necessary for the translation initiation, but the possibility remains that either the upstream UTR from the first ATG codon has been truncated or initiation from the internal ATG sites of the truncated or 5′-intron-retaining clones was induced in the *in vitro* system.[6]

3.2. *Functional classification of the predicted gene products*

The gene products newly predicted were tentatively sorted into the following six functional categories (1 to 6), and those of unclassified (7), and no clue (8) on the basis of similarity to known proteins or protein motifs linked to certain functions:

(1) Proteins relating to cell signaling and cell-cell communication such as protein kinases, adhesion molecules, receptors, and intracellular signal transducers (cell signaling/communication).
(2) Proteins relating to nucleic acid synthesis and its regulation such as transcription factors, helicases and splicing factors (nucleic acid managing).
(3) Proteins relating to protein synthesis including modification and degradation enzymes (protein managing).
(4) Proteins relating to cell structure and motility including cytoskeletons, membrane skeletons, extracellular matrixes, motors (cell structure/motility).
(5) Proteins relating to cell division (cell division).
(6) Proteins relating to metabolisms (metabolism).
(7) Proteins not classified into above functional categories and hypothetical proteins registered in databases (unclassified).
(8) Hypothetical proteins with no clue for functional classification (no clue).

The results of analysis are summarized in Table 2, which lists the name of database files which gave the highest score in the similarity queries, the identity of amino acid sequences and the sequence length from which the identity was deduced. The overall degree of similarity was also indicated by dividing them into nearly "identical" (> 90% to entries of human proteins), "homologous" (> 90% to non-human protein entries), "related" (30–90% to any entries) and "weakly related" classes (< 30% to any entries). Note that the identical class includes seven clones whose sequences were reported during the preparation of this paper. The sequence comparison of these clones with the reported proteins suggests that some of the products were originated by alternative splicing.

The distribution of the predicted gene products in the six functional classes seems to be essentially in agreement with that of the entries of large proteins (> 1000 amino acid residues) in the database of human proteins. Other sequence features noteworthy are summarized below.

1. Eleven out of 21 genes which were classified into category 2 (nucleic acid managing) coded for DNA-binding proteins carrying zinc finger motifs.

2. Only one gene (KIAA0361) was assigned to category 6 (metabolism). Furthermore, this gene product showed similarity to prokaryotic protein entries. The rare occurrence of metabolic enzymes is consistent with the search data of the large protein entries in the database.[2]

3. The occurrence of receptors, channels and adhesion proteins in the gene products assigned to category 1 (cell signaling/communication) appeared to be lower than that expected from the large protein entries

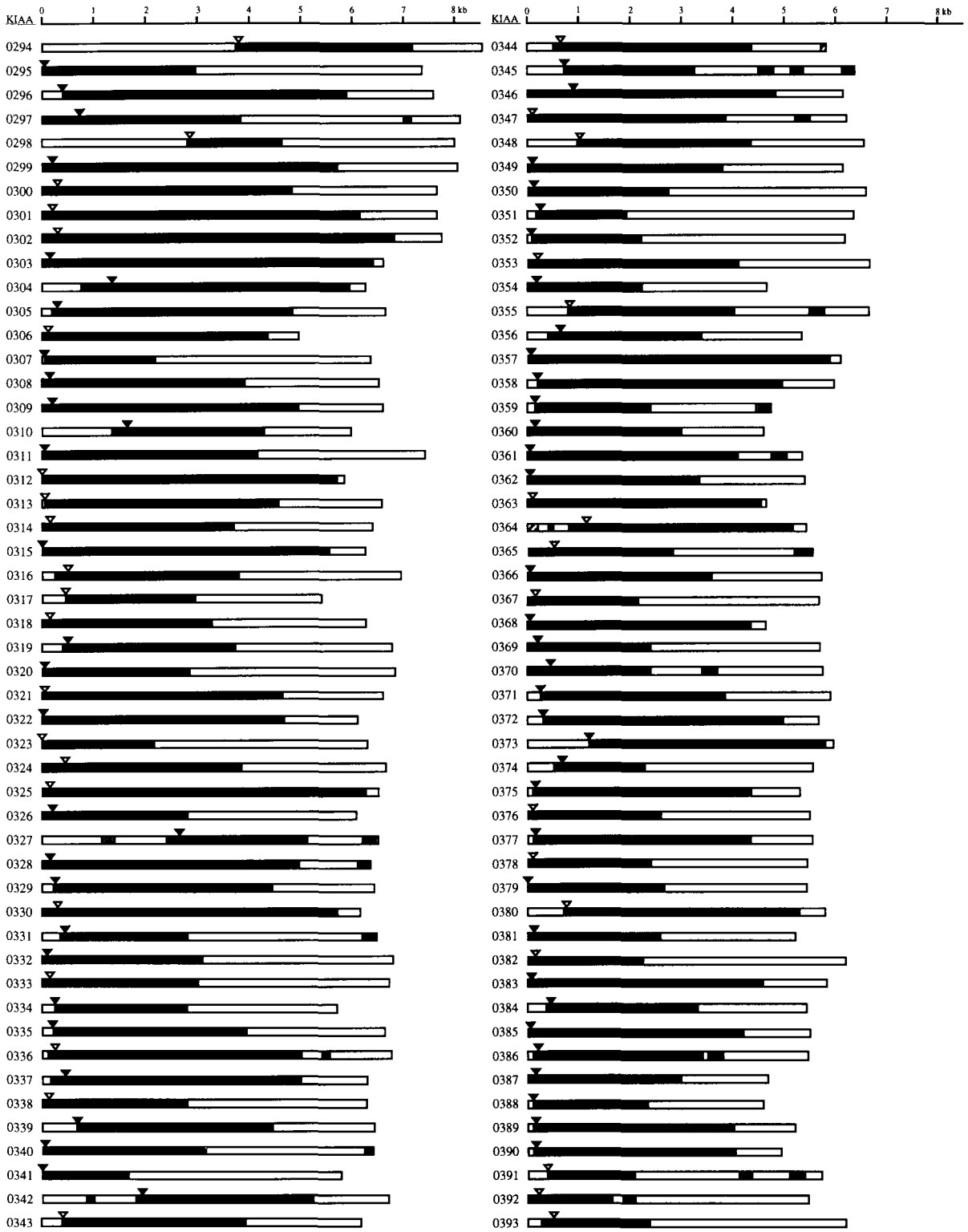

Figure 1. Physical maps of the 100 cDNA clones analyzed. The horizontal scale represents the cDNA length in kilobases, and the gene numbers corresponding to respective cDNAs are given on the left. The ORFs and untranslated regions are shown by solid and open boxes, respectively. The positions of the first ATG codons in the cDNAs are indicated by triangles and those in the contexts of Kozak's rule are illustrated by solid triangles. Alu sequences and other repetitive sequences are represented 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)	Apparent molecular mass (kDa)[c]	Chromosomal location[d]
294	AB002292	8,467	1,121	>100	8
295	AB002293	7,326	981	>100	15
296	AB002294	7,604	1,829	>100	16
297	AB002295	8,039	1,271	>100	14
298	AB002296	8,001	597	87	11
299	AB002297	8,063	1,907	>100	3
300	AB002298	7,621	1,608	>100	5
301	AB002299	7,651	2,047	>100	6
302	AB002300	7,762	2,276	>100	11
303	AB002301	6,629	2,137	>100	5
304	AB002302	6,252	1,529	>100	19
305	AB002303	6,632	1,539	>100	5
306	AB002304	4,964	1,451	>100	19[e]
307	AB002305	6,415	706	83	15
308	AB002306	6,452	1,297	>100	16
309	AB002307	6,648	1,668	>100	16
310	AB002308	5,955	881	>100	9
311	AB002309	7,418	1,386	>100	14
312	AB002310	5,842	1,906	>100	X
313	AB002311	6,568	1,499	>100	4
314	AB002312	6,424	1,240	>100	12
315	AB002313	6,264	1,845	>100	22
316	AB002314	6,935	1,094	>100	13
317	AB002315	5,402	823	97	14
318	AB002316	6,322	1,106	>100	12
319	AB002317	6,791	1,072	>100	6
320	AB002318	6,865	949	>100	15
321	AB002319	6,540	1,542	>100	14
322	AB002320	6,102	1,562	>100	7
323	AB002321	6,227	724	92	14
324	AB002322	6,619	1,288	88	16
325	AB002323	6,494	2,087	>100	14
326	AB002324	6,045	927	>100	16
327	AB002325	6,486	820	93	5
328	AB002326	6,411	1,661	>100	2
329	AB002327	6,403	1,411	>100	14
330	AB002328	6,143	1,902	>100	22
331	AB002329	6,474	775	95	7
332	AB002330	6,823	1,028	>100	3
333	AB002331	6,692	991	>100	20
334	AB002332	5,715	846	>100	4
335	AB002333	6,639	1,263	95	10
336	AB002334	6,773	1,583	>100	2
337	AB002335	6,289	1,510	>100	11
338	AB002336	6,263	934	>100	9
339	AB002337	6,446	1,260	>100	16
340	AB002338	6,395	1,053	>100	6
341	AB002339	5,721	546	82	5
342	AB002340	6,691	1,098	94	3
343	AB002341	6,218	1,180	>100	7
344	AB002342	5,799	1,246	>100	12
345	AB002343	6,387	842	84	5
346	AB002344	6,133	1,620	>100	17
347	AB002345	6,218	1,246	>100	6[e]
348	AB002346	6,562	1,113	>100	6
349	AB002347	6,158	1,275	>100	6
350	AB002348	6,607	917	>100	16
351	AB002349	6,336	557	64	9
352	AB002350	6,170	712	87	12
353	AB002351	6,651	1,374	>100	15
354	AB002352	4,631	677	>100	9
355	AB002353	6,657	1,070	>100	19
356	AB002354	5,371	926	>100	17
357	AB002355	6,106	1,961	>100	17
358	AB002356	5,942	1,581	>100	11
359	AB002357	4,724	747	>100	20
360	AB002358	4,611	1,001	>100	14
361	AB002359	5,338	1,371	>100	17
362	AB002360	5,391	1,108	>100	13
363	AB002361	4,642	1,522	>100	7
364	AB002362	5,413	1,327	>100	X
365	AB002363	5,475	939	>100	19
366	AB002364	5,774	1,201	>100	4
367	AB002365	5,654	716	>100	9
368	AB002366	5,665	1,441	>100	9
369	AB002367	5,703	729	93	13
370	AB002368	5,724	801	84	16
371	AB002369	5,886	1,198	>100	20
372	AB002370	5,716	1,564	>100	5
373	AB002371	5,967	1,539	>100	12
374	AB002372	5,530	538	72	20
375	AB002373	5,324	1,404	>100	9
376	AB002374	5,527	889	>100	22
377	AB002375	5,556	1,406	>100	15
378	AB002376	5,434	808	>100	3
379	AB002377	5,457	882	>100	3
380	AB002378	5,790	1,522	>100	1
381	AB002379	5,432	864	>100	6
382	AB002380	6,203	750	>100	11
383	AB002381	5,810	1,520	81	10
384	AB002382	5,423	939	>100	11
385	AB002383	5,492	1,370	>100	X
386	AB002384	5,471	1,068	>100	6
387	AB002385	4,679	1,006	>100	7
388	AB002386	4,606	747	>100	17
389	AB002387	5,212	1,285	>100	6
390	AB002388	4,935	1,300	>100	19
391	AB002389	5,677	567	70	14
392	AB002390	5,435	554	64	8
393	AB002391	6,067	618	72	15

[a] Accession numbers of DDBJ, EMBL and GenBank databases.
[b] Values excluding poly(A) sequences.
[c] Approximate molecular masses estimated by SDS-PAGE.
[d] Chromosome numbers identified by using GeneBridge 4 radiation hybrid panel unless specified.
[e] Chromosome numbers determined by using CCR human-rodent hybrid panel.

in the database. This may be due to the fact that such molecules of biological interest have been preferentially investigated to date. On the other hand, this study allowed us to identify seven gene products which are involved in the G-protein-mediated signal transduction system. This value is considerably high, compared with that in the large protein entries in the database.

The complete sequence catalog of the protein components in budding yeast is now available (non-redundant 6023 yeast protein sequences retrieved via ftp at ftp://genome-ftp.stanford.edu/pub/yeast/yeast_ORFs). Thus, the protein sequences predicted in this paper were compared with those in this database, and only 4 genes (KIAA0325, 0357, 0361 and 0389) were found to have significant similarities to yeast proteins over the 50% region of the entire query sequences. As listed in Table 2, three of them were motor proteins, dynein (KIAA0325 and 0357) and myosin (KIAA0389), which are known to be highly conserved during the evolution of eukaryotic organisms, and the remaining one (KIAA0361) was the metabolic enzyme mentioned previously. We noted that many of the gene products predicted in this paper carried some sequences which showed significant similarities to yeast proteins but the lengths of the similarity regions were short relative to the query sequences. As far as the protein sequences reported here are concerned, the counterparts of at least 96% of the human gene products appear to be absent in the unicellular eukaryotic organism.

3.3. Expression profiles of the predicted genes

The steady-state levels of individual mRNAs in 14 different tissues were analyzed by the RT-PCR method. To examine the fidelity of this method, we compared the expression patterns using three representative genes, which exhibited ubiquitous, tissue-specific and very low expression patterns by Northern analysis (lower panels in Fig. 2A, B, and C). As shown in the upper panels of Fig. 2A, B, and C, the RT-PCR method produces expression patterns comparable to those by Northern analysis. The result also demonstrated that the sensitivity of the former method is higher than that of the latter under the conditions employed. The possibility of generating PCR products from contaminated genomic DNA in the original poly $(A)^+$ RNA preparations was excluded by PCR with a primer set from split exons and also by an experiment in which the reverse transcription step was omitted.

The quantitative measurement of mRNA levels by RT-PCR, however, requires careful control experiments for each gene, because the quantifiable range of RT-PCR is relatively narrow, and the efficiency of amplification is significantly influenced by the sequences of the primer

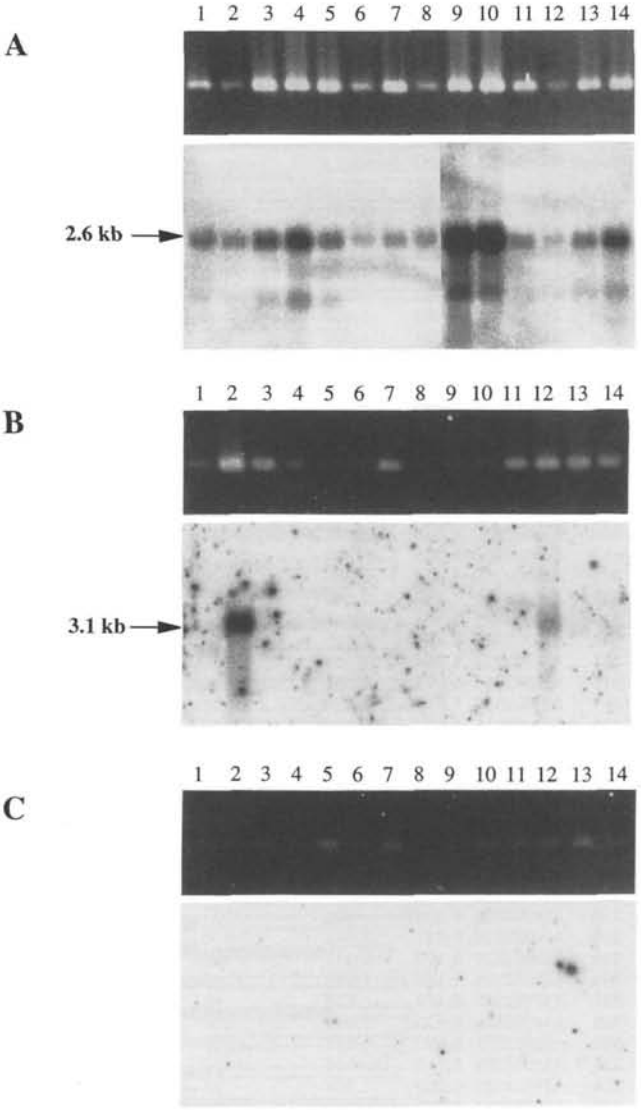

Figure 2. Comparison of gene expression profiles analyzed by RT-PCR and Northern analysis. The expression levels of three representative genes which show ubiquitous (A: KIAA0085[8]), tissue-specific (B: KIAA0273[9]) and very weak expression (C: KIAA0276[9]) were analyzed by RT-PCR (upper in each panel) and Northern blotting (lower in each panel). Lane 1, heart;2, brain; 3, placenta; 4, lung; 5, liver; 6, skeletal muscle; 7, kidney; 8, pancreas; 9, spleen; 10, thymus; 11, prostate; 12, testis; 13, ovary; 14, small intestine. For Northern analysis, cDNA fragments were randomly labeled as probes and hybridization was carried out as described previously.[1] Human MTN blots were purchased from CLONTECH laboratories, Inc. (USA).

set used.[7] For practical purposes, in this study, we used the RT-PCR analysis as a rough means of distinguishing the tissues which express a gene of interest from those in which it remains dormant, and the band intensities generated from varying amounts of the cDNA clone carrying this gene are indicated in Fig. 3, lanes 1–5, for compari-

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/communicatiot	0294	W	ABR protein 2 (H)	B47485	27.2	213
	0299	R	KIAA0209 (H)	D86964	31.2	1603
	0300	R	putative interleukin-16 precursor (H)	S81601	36.7	313
	0303	R	S/T protein kinase MAST205 (M)	A54602	43.9	1146
	0311	I	A-kinase anchor protein AKAP100 (H)	U17195	99.8	654
	0313	W	KIAA0277 (H)	D87467	32.6	313
	0315	R	SEP (H)	X87904	53.9	802
	0321	W	Hrs (M)	D50050	31.1	103
	0327	R	protocadherin 43 (H)	L11373	42.1	770
	0332	W	homolog of Drosophila splicing regulator (H)	U08377	19.6	409
	0337	W	breakpoint cluster region BCR (H)	U07000	27.0	330
	0340	W	rabphilin-3A (M)	JX0338	29.0	124
	0343	I	hBRAVO/Nr-CAM precursor (H)	U55258	97.1	803
	0344	W	calphotin (D)	A47283	19.9	710
	0345	R	protocadherin 43 (H)	L11373	44.3	705
	0347	W	period clock protein homolog (Pa)	U12772	38.7	155
	0348	R	synaptojanin (R)	U45479	42.4	932
	0350	R	hypothetical 100.9 kD protein C34E10.3 (Ce)[g]	P46578	31.4	810
	0351	W	Ras-GRF2 (M)	U67326	31.5	241
	0362	R	Ost oncogene (R)	S51620	87.5	872
	0364	R	alpha-1B-glycoprotein (H)	P04217	26.6	447
	0369	R	Ca2+/calmodulin-dependent protein kinase (R)	S50193	43.8	292
	0371	R	myotubularin MTM1 (H)	U46024	47.4	270
	0380	R	Lsc oncogene (M)	U58203	42.5	534
	0382	R	Lsc oncogene (M)	U58203	48.0	356
	0384	H	p120 protein (M)	P30999	96.5	882
	0387	I	phogrin (H)	U66702	99.7	988
Nucleic acid managing	0295	W	zinc finger homeodomain protein ATBF1-A (H)[f]	L32832	31.9	47
	0296	W	zinc finger protein XFDL 156 (X)	U65898	23.2	413
	0304	R	zinc finger protein HRX (M)	P55200	36.8	883
	0306	W	HMG-box transcription factor SOX-18 (M)	L35032	25.5	145
	0307	H	Arnt2 (R)	U61157	96.5	706
	0309	W	218kD Mi-2 (H)	X86691	32.4	367
	0312	H	DNA binding protein URE-B1(R)	P51593	91.6	308
	0314	W	p300/CBP-associated factor P/CAF (H)	U57317	37.3	126
	0324	R	clone 1/6 nuclear protein (Em)	L41834	34.4	381
	0326	R	zinc finger protein ZFP-35 (M)	P15620	55.9	451
	0333	R	KIAA0244 (H)	D87685	50.0	186
	0334	W	sim transcription factor (M)	U42554	21.0	453
	0335	W	zinc finger Y-chromosomal protein1(X)	Q01611	25.0	248
	0352	W	zinc finger 5 protein ZF5 (C)	U51640	33.1	118
	0354	W	Z13 protein (M)	S59069	23.1	350
	0360	R	zinc finger protein Xsal-1 (X)	L46583	32.7	686
	0363	R	transcriptional activator alpha-NAC Naca (M)	U48364	24.9	1098
	0383	R	zinc finger protein MOZ (H)	U47742	28.3	1265
	0388	I	enhancer of zeste homolog 1 Ezh1 (H)	U50315	99.9	747
	0390	W	KIAA0222 (H)	D86975	32.6	631
Cell structure/motility	0301	W	probable membrane protein YLR106c (Sc)	S64942	26.2	866
	0302	R	spectrin beta-G chain	A44159	63.0	2201
	0310	W	alpha-5 collagen type IV COL4A5 (H)	M58526	21.9	351
	0319	W	twitchin (Ce)	S07571	15.8	437
	0320	R	talin (M)	P26039	76.8	950
	0323	W	nonfibrillar collagen (Su)	S64572	25.2	218
	0325	H	dynein heavy chain (R)	I58139	98.7	2087
	0330	W	procollagen alpha (M)	P02463	20.1	149
	0331	R	semaphorin H (M)	Z80941	87.4	775
	0336	W	myosin heavy chain	U32574	18.0	1108
	0338	R	PROTEIN 4.1 (BAND 4.1) (H)	P11171	45.7	670

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 structure/motility	0341	W	desmoplakin I (H)	M77830	21.9	233
	0353	W	nestin (R)	P21263	15.5	1345
	0357	R	dynein beta chain (Su)	P23098	71.3	1524
	0359	H	kinesin KIF3B (M)	A57107	98.0	747
	0376	W	CENP-E (H)	Q02224	17.7	666
	0378	W	novel golgi-associated protein GCP360 (R)	D25543	19.5	688
	0379	W	ankyrin 3 (H)	A55575	30.8	558
	0389	H	unconventional myosin VI (M)	U49739	89.3	1055
Cell division	0329	W	sepB protein (An)	S54152	21.9	114
	0367	R	E1B 19K/Bcl-2-interacting protein NIP2 (H)	I38864	53.8	210
	0373	W	mitosin (H)	U30872	17.6	1353
	0381	W	diaphanous protein (D)	P48608	25.8	881
Protein managing	0317	W	ubiquitin ligase Nedd4 (R)	U50842	36.3	394
	0322	R	ubiquitin ligase Nedd4 (R)	U50842	47.8	400
	0349	W	N-end-recognizing protein (Sc)	P19812	19.9	710
Metabolism	0361	R	FGAM synthase (D)	P35421	49.7	1365
Unclassified	0305	W	Caenorhabditis elegans cosmid D1022 (Ce)	U23517	17.6	705
	0346	W	Caenorhabditis elegans cosmid D2021 (Ce)	U23513	40.9	494
	0356	W	KIAA0226 (H)	D86979	31.0	248
	0358	I	DENN (H)	U44953	91.9	975
	0368	R	Caenorhabditis elegans cosmid D2045 (Ce)	Z35639	33.9	826
	0372	W	superkiller 3 protein (Sc)	P17883	25.0	244
	0374	W	PUFF II/9-2 protein precursor (Fg)	P22312	23.2	125
	0377	R	hypothetical protein YLR410w (Sc)	S59376	45.2	283
	0385	I	DXS6673E (H)	X95808	99.1	1370
	0386	I	Diff48 (H)	U49187	99.0	308
	0392	W	Caenorhabditis elegans cosmid R07E4 (Ce)	U39652	33.1	514
No clue	0297		none			
	0298		none			
	0308		none			
	0316		none			
	0318		none			
	0328		none			
	0339		none			
	0342		none			
	0355		none			
	0365		none			
	0366		none			
	0370		none			
	0375		none			
	0391		none			
	0393		none			

a) Classifications based on the annotations of their homologous protein entries in the databases unless otherwise stated.
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%). The gene products in class I (> 90%) include alternative splicing products of reported genes.
c) Organisms in which these entries were identified are given in parentheses: An, *Aspergillus nidulans*; C, chicken; Ce, *Caenorhabditis elegance*; *D, Drosophila melanogaster*; *Em, Ensis minor*; Fg, *Fungus gnat*; H, human; M, mouse; Pa, *Periplaneta americana*; R, *rat*; Sc, *Saccharomyces cerevisiae*; Su, *sea urchin*; X, *Xenopus leavis*.
d) Accession numbers of homologous entries in DDBJ/EMBL/GenBank/OWL/SWISS-PLOT/PIR database are shown.
e) The values were obtained by the FASTA program.
f) Classifications based on the sequence motifs.

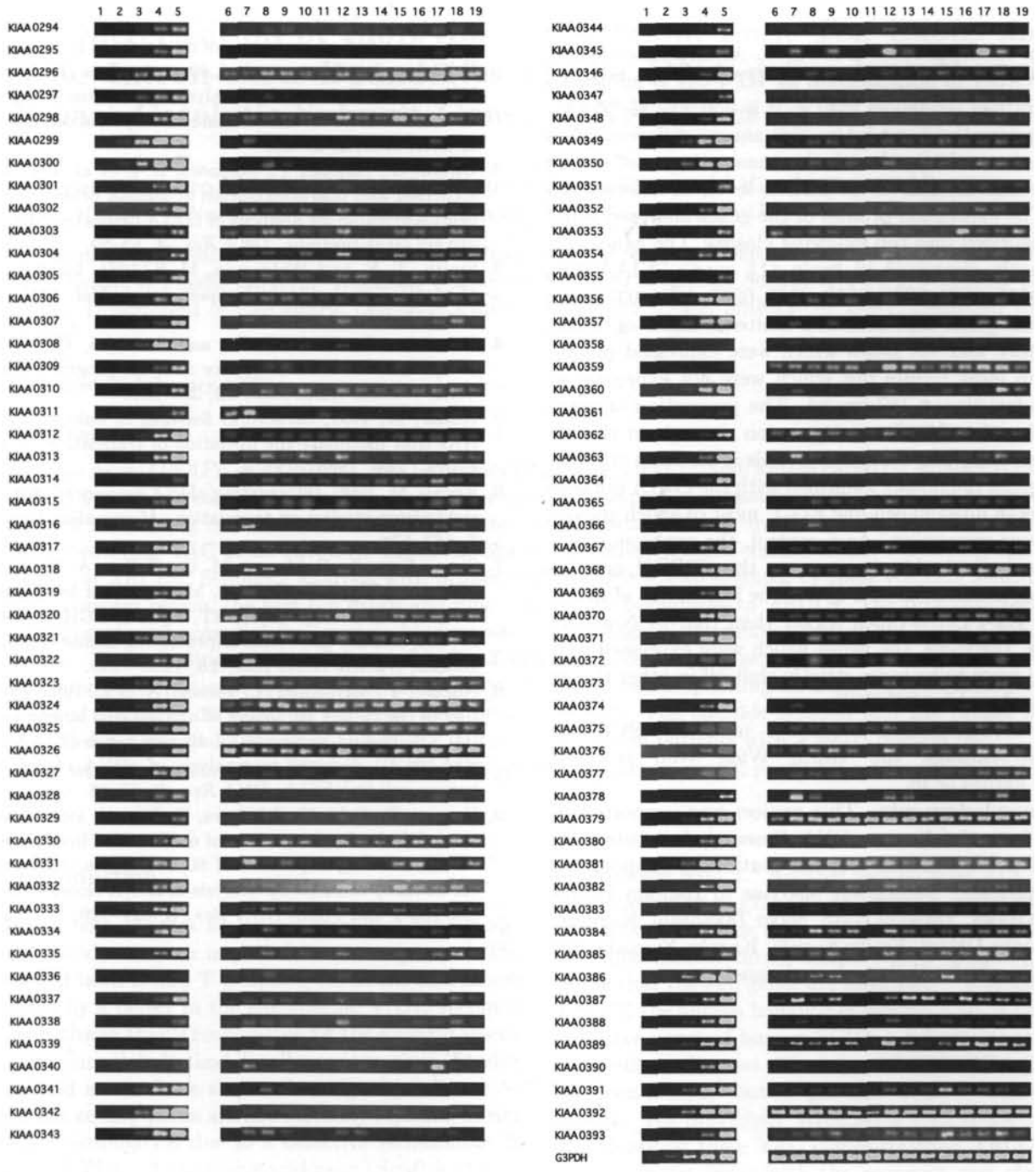

Figure 3. Expression profiles of 100 newly identified genes in 14 different tissues examined by RT-PCR. Electrophoretically resolved bands of the PCR products for individual genes are shown. Gene numbers are given on the left and G3PDH gene expression was analyzed as a positive control. In each set of electrophoretic patterns, lanes 1 to 5 show the PCR products derived from serial tenfold dilutions of a cDNA clone of interest (from 0.1 fg to 1 pg, respectively) for the estimation of the PCR amplification efficiency. Lanes 6 to 19 are electrophoretic patterns of the PCR products originated from mRNAs of 14 different tissues: lane 6, heart; lane 7, brain; lane 8, placenta; lane 9, lung; lane 10, liver; lane 11, skeletal muscle; lane 12, kidney; lane 13, pancreas; lane 14, spleen; lane 15, thymus; lane 16, prostate; lane 17, testis; lane 18, ovary; lane 19, small intestine. All the PCR products were run on 2.5% NuSeive GTG agarose gel and detected by staining with ethidium bromide. Several cDNA clones appear to show no bands in any lane, but a faint band is seen in at least one of the tissues in the original photograph.

son of the efficiency of PCR with the primer set for each gene.

The expression profiles of individual transcripts in 14 different tissues are shown in Fig. 3. As mentioned above, the efficiency of amplification by RT-PCR is influenced by the primer sequences used — it is not possible to simply compare the band intensities among different genes — but those obtained with the same primer set should provide information on the mRNA levels in different tissues. The expression profiles of the genes analyzed could be categorized into the following classes: The genes predominantly expressed in brain (11 genes: KIAA0299, 0311. 0316, 0319, 0322, 0358, 0363, 0369, 0374, 0378, and 0390), the genes expressed in a limited number of tissues (31 genes), and the genes which were expressed ubiquitously in most tissues but which were not expressed at all in a few tissues (47 genes). The proportion of genes exhibiting the ubiquitous expression throughout the tissues was fairly low (11 genes). This is in sharp contrast to the genes previously identified with the cDNA libraries of a human myeloid cell line KG-1, most of which showed ubiquitous expression. As a control, the expression profile of a typical ubiquitous gene, the G3PDH gene, is shown at the bottom of Fig. 3. The sensitivity of detection by RT-PCR is much higher than that by Northern analysis; therefore, the genes which were expressed only in brain seem to be more strictly shut off in other tissues when analyzed by RT-PCR.

The actual primer sequences used for PCR are available through the World Wide Web at http://www.kazusa.or.jp.

Acknowledgments: This project was supported by grants from the Kazusa DNA Research Institute. We thank Dr. M. Takanami for his continuous support and encouragement. Thanks are also due to Tomomi Tajino, Keishi Ozawa, Tomomi Kato, Seiko Takahashi, Kazuhiro Sato, Akiko Ukigai, Emiko Suzuki, Kazuko Yamada, and Naoko Suzuki for their technical assistance.

References

1. 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-KIAA0040) deduced by analysis of randomly sampled cDNA clones from human immature myeloid cell line KG-1, *DNA Res.*, **1**, 27–35.
2. 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.
3. Smith, T. F. and Waterman, M. S. 1981, Identification of common molecular subsequences, *J. Mol. Biol.*, **147**, 195–197.
4. Devereux, J., Haeberli, P., and Smithies, O. 1984, A comprehensive set of sequence analysis programs for the VAX, *Nucleic Acids Res.*, **12**, 387–395.
5. Kozak, M. 1991, Structural features in eukaryotic mRNAs that modulate the initiation of translation, *J. Biol. Chem.*, **266**, 19867–19870.
6. Kozak, M. 1996, Interpreting cDNA sequences: some insights from studies on translation, *Mammalian Genome*, **7**, 563–574.
7. Singer-Sam, J., Robinson, M. O., Bellve, A. R., Simon, M. I., and Riggs, A. D. 1990, Measurement by quantitative PCR of changes in HPRT, PGK-1, PGK-2, ARPT, MTase, and Zfy gene transcripts during mouse spermatogenesis, *Nucleic Acids Res.*, **18**, 1255–1259.
8. Nagase, T., Miyajima, N., Tanaka, A. et al. 1995, Prediction of the coding sequences of unidentified human genes. III. The coding sequences of 40 new genes (KIAA0081-KIAA0120) deduced by analysis of cDNA clones from human cell line KG-1, *DNA Res.*, **2**, 37–43.
9. Nagase, T., Seki, N., Ishikawa, K.-I. et al. 1996, Prediction of the coding sequences of unidentified human genes. VI. The coding sequences of 80 new genes (KIAA0201-KIAA0280) deduced by analysis of cDNA clones from cell line KG-1 and brain, *DNA Res.*, **3**, 321–329.