DNA RESEARCH 4, 53–59 (1997)

Construction and Characterization of Human Brain cDNA Libraries Suitable for Analysis of cDNA Clones Encoding Relatively Large Proteins

Osamu OHARA,* Takahiro NAGASE, Ken-ichi ISHIKAWA, Daisuke NAKAJIMA, Miki OHIRA, Naohiko SEKI, and Nobuo NOMURA

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

(Received 20 January 1997)

Abstract

Analysis of proteins registered in the PIR protein database implied that most of relatively large proteins are related to important functions in higher multicellular organisms, but not many large proteins have been registered to date. To establish a protocol for efficient analysis of cDNA clones coding for large proteins, we constructed a series of strictly size-fractionated cDNA libraries of human brain, where the average insert sizes of cDNA clones ranged from 3.3 kb to 10 kb. As judged by hybridization analysis with probes derived from mRNAs of known sizes, the libraries with insert sizes up to 7 kb, at least, contained the clones corresponding to full-length transcripts in addition to truncated products of longer transcripts, but few chimeric clones. Using one of the fractionated libraries with an average insert size of 7 kb, the single-pass sequences from both the ends of randomly sampled clones were determined and searched against DNA databases. Approximately 90% of the clones were found to be new with respect to their 5′-sequences, while their 3′-sequences were frequently similar to the registered expression sequence tags. Examination of the protein-coding capacity in an *in vitro* transcription/translation system showed that about 20% of the clones direct the synthesis of proteins with apparent molecular masses larger than 50 kDa. The set of libraries constructed here should be very useful for the accumulation of sequence data on large proteins in the human brain.

Key words: large proteins; cDNA library; size fractionation; single-pass sequencing

1. Introduction

Tremendous efforts have been put into the characterization of human cDNAs and more than 600,000 single-pass cDNA sequences (expression sequence tags: ESTs) have been deposited in the public DNA databases to date.[1–3] Although this approach has proved to be powerful for tagging and mapping transcript complements on the genome, it is doubtful that the complete protein sequences can be assembled from the EST data, as most ESTs fall in the region about 2 kb from the poly(A) tails of cDNA.[4] Based on the view that the accumulation of complete protein-coding sequences is essential for investigation of biological functions of gene products, we have implemented a project for sequencing human cDNAs which correspond to relatively long and nearly full-length transcripts. In this project, human myeloid cell line KG-1 was mainly used as the source of cDNAs and so far, the sequences of 280 clones carrying the coding regions of unidentified human genes have been reported.[5] The average size of the clones analyzed in this project was 4.4 kb. Analysis of potential open reading frames (ORFs) indicated that the ORF length of cDNA increases with the size of the cDNA, and thus we could eventually accumulate the information on human proteins with relatively large sizes through this project.

On the other hand, analysis of proteins registered in the public databases demonstrated that most of the relatively large proteins were related to important functions in higher multicellular organisms. However, not many large proteins have been reported to date, if the proteins longer than 1,000 amino acid residues are taken, only 460 of the 9213 human protein entries in the PIR database (release no. 50). To efficiently accumulate the predicted sequence data on large proteins, we have established a protocol for analysis of cDNA clones encoding large proteins.

The source of cDNA clones used was human brain, because its mRNA pool is assumed to have greater com-

Communicated by Mituru Takanami

* To whom correspondence should be addressed. Tel. +81-438-52-3913/3915, Fax. +81-438-52-3914, E-mail:ohara@kazusa.or.jp

plexity and there must be many biologically important large proteins. According to our sequence data on human cDNA clones, the minimum size of cDNAs coding for large proteins should be about 4 kb. Thus, we strictly fractionated the brain cDNA libraries in order of increasing insert sizes, and purified them by repeating the fractionation step. Then, the contents of full-length cDNA clones and of non-chimeric clones in fractions were examined, and the libraries with insert sizes of 4 kb to 7 kb were found to be suitable for analysis of cDNA clones encoding large proteins. With one of the libraries having the average insert size of 7.0 kb, it was shown that the library contains unidentified cDNA clones at a higher frequency, and that approximately 20% of the clones can direct the synthesis of proteins with apparent molecular masses larger than 50 kDa in an *in vitro* transcription/translation system.

2. Materials and Methods

2.1. *cDNA library construction*

Double-stranded cDNA fragments were generated by the use of reverse transcriptase (SuperScript II, GibcoBRL, USA), a $(dT)_{15}$ primer carrying the *Not* I site at the 5′-moiety (5′-pGACTAGTTCTAGATCG CGAGCGGCCGCCC$(T)_{15}$-3′) and poly(A)$^+$ RNA of human whole brain (Clontech, USA) according to the supplier's instructions. After ligation of the *Sal* I adaptor followed by *Not* I digestion, the resultant cDNAs were electrophoresed on 1% low-melting agarose gel to remove cDNA fragments smaller than 3 kb. The cDNA fragments larger than 3 kb were then ligated with the *Sal* I-*Not* I-digested pBluescript IISK+ vector and introduced into *Escherichia coli* cells by electroporation (ElectroMax DH10B cells, GibcoBRL, USA). Plasmids were extracted from approximately 8×10^6 independent ampicillin-resistant colonies grown on agar plates by the standard alkaline/sodium dodecylsulfate method. The recovered plasmids, which were mostly in a covalently closed circular form, were resolved by agarose gel electrophoresis into ten fractions in order of increasing insert cDNA size (larger than 3 kb). The plasmids in each fraction were re-introduced to *E. coli* cells and recovered from colonies (more than 10^6 per fraction) on agar plates as described above except that the transformed cells were grown for 3 hr at 37°C in LB medium containing ampicillin (100 μg/ml) with shaking. The extracted plasmids were electrophoresed and those corresponding to the original fractionated size were retrieved. This size selection step was repeated at least twice for fractions up to 8 kb and three times for those larger than 8 kb.

2.2. *Hybridization analysis of size-fractionated plasmid cDNAs*

The plasmids constituting the respective size-fractionated libraries were digested with *Not* I and *Mlu* I, and run on a 0.7% agarose gel. The resolved DNA fragments were transferred onto a nylon membrane after alkaline denaturation/neutralization by standard methods. The cDNA fragment (1.3 kb) of the human neurodap1 gene was prepared by polymerase chain reaction (PCR) with following two primers: 5′-GACAGACAGAACATTCACCTG-3′ and 5′-TCTCATTTCAACTGTCAAGG-3′.[6] In the same way, the cDNA fragment (1.5 kb) of a human gene containing the tetratricopeptide repeat motifs on the Down syndrome region (TPRD gene) was obtained by the polymerase chain reaction with primers, 5′-AAACAAGATCTCAA AGACGG-3′ and 5′-TGGTGGATGACACTTTAGTG-3′.[7] These two cDNA fragments were labeled with fluorescein using the Gene Images labeling system (Amersham, UK) and used as probes. The α-spectrin cDNA probe (6.0 kb) was prepared from a partial cDNA clone encodng non-erythroid α-spectrin[8] and labeled with [^{32}P]dCTP using the RadPrime^{TM} labeling kit from GibcoBRL (USA). The hybridization was conducted at 65°C as described by Church and Gilbert.[9] After the hybridization, the membrane was washed with 2 × SSC (1 × SSC = 0.15 M NaCl and 15 mM sodium citrate, pH 7.0) containing 1% sodium dodecylsulfate (SDS) at 65°C for 30 min, and then twice with 0.1 × SSC containing 0.1% SDS at 65°C for 30 min. The hybridization signals were detected with Gene Images CDP-Star detection module (Amersham, UK; for the fluorescein-labeled probes) or with the BAS-2000 imaging system (Fuji film, Japan; for the ^{32}P-labeled probe).

2.3. *DNA sequencing*

Plasmid DNAs were routinely prepared with robots (PI-100, Kurabo, Japan), and dye-primer cycle sequencing reactions were performed using ABI PRISM^{TM} cycle sequencing kits (Perkin-Elmer Cetus, USA) and the robotic reaction system (CATALYST Turbo, Perkin-Elmer Cetus, USA). The resultant products were analyzed by ABI 373A or 377 DNA sequencers with an ABI sequence analysis system, INHERIT. Dye-labeled M13 forward and reverse primers were used for single-pass sequencing from the ends and also for the entire sequencing of cDNA inserts by the shotgun strategy. For deduction of the entire sequence, a cDNA insert was excised from the vector and purified by agarose gel electrophoresis. The recovered cDNA insert was self-ligated, dephosphorylated with calf intestinal alkaline phosphatase, and then sheared by sonication. The resulting fragments of 700–1000 bp were blunt-ended, retrieved by agarose gel

electrophoresis and then ligated with dephosphorylated, *Sma* I-digested M13mp18 vector. We routinely analyzed 16 randomly isolated shotgun clones per 1 kb cDNA inserts. Gaps that remained were filled by dye-terminator or dye-primer sequencing using appropriate primers of PCR products designed for gap filling. The entire sequences of cDNA clones were constructed from the data on both the strands. Homology search and other analyses of the obtained sequences were performed with the GCG software package.[10]

2.4. *In vitro transcription/translation assay*

Plasmids treated with RNase A, passed through ADVAMAX beads (AGTC Inc., USA) for removal of RNase A, and subjected to an *in vitro* transcription/translation system (TNT T7 coupled reticulocyte lysate system, Promega Co., USA) in the presence of [^{35}S]methionine (Amersham, UK). The products were resolved on 10% polyacrylamide gel containing 0.1% SDS as described by Laemmli[11] and detected with a BAS-2000 image analyzer (Fuji film, Japan).

2.5. *Radiation hybrid mapping*

Single-pass end sequences were mapped along the human genome by using GeneBridge 4 radiation hybrid panel (Research Genetics, Inc., USA). Detailed experimental conditions for the radiation hybrid mapping of respective sequences will be described elsewhere (manuscript in preparation). Software for analysis of results from the panel was obtained via the World Wide Web at http://www.sph.umich.edu/group/statgen/software.

3. Results and Discussion

3.1. *Biological significance of large proteins*

The entire genomic sequences of four bacteria strains (*Haemophilus influenzae* Rd,[12] *Mycoplasma genitalium*,[13] *Cyanobacterium synechosystis* sp. PCC6803,[14] and *Methonococcus jannaschii*[15]) and *Saccharomyces cerevisiae* (available via the World Wide Web at http: //speedy.mips.biochem.mpg.de/mips /yeast/) have been deduced and the potential coding regions along the genomes have been predicted and deposited in the PIR protein database. When the size distribution of the whole proteins encoded in the respective genomes were compared, we noted that the distribution of yeast proteins is distinctive from those of the four prokaryotic organisms. As shown in Fig. 1, the four prokaryotic organisms exhibit almost identical distribution patterns in spite of the divergence of their genome size (0.58 Mb to 3.6 Mb), and 97–100% of proteins were smaller than 1,000 amino acid residues. In contrast, the

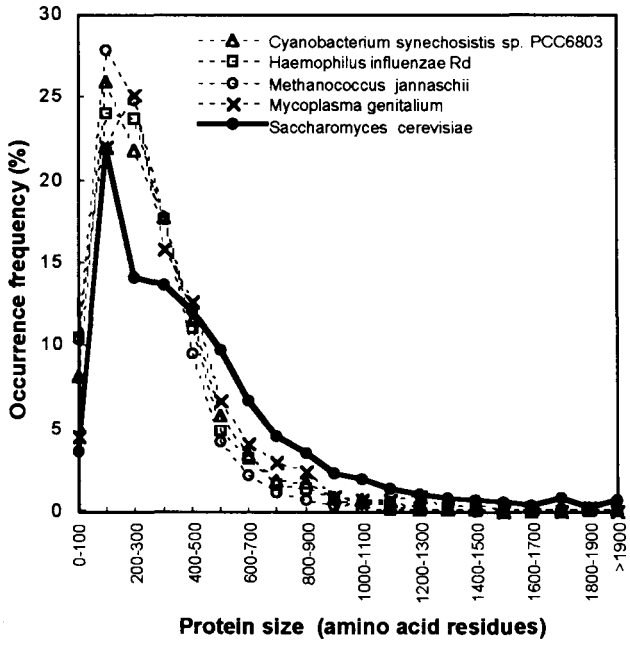

Figure 1. Comparison of the size distributions of protein complements in four prokaryotes and yeast which were predicted from their genomic sequences.
The predicted protein sizes in respective organisms were compiled from the PIR database (release no. 50). The frequencies of occurrence were normalized and given in percentages.

size distribution of yeast proteins shows a broad shoulder toward the longer side, and about 10% of all proteins were larger than 1,000 amino acid residues. The result suggests that the presence of proteins with large molecular weights is a characteristic of eukaryotic organisms.

We then surveyed the large proteins registered in the PIR protein database (release no. 50). Taking into account the proteins larger than 1,000 amino acid residues, the entry numbers are 459 for human (total entry number, 9213), 202 for rat (total entry number, 3740), and 224 for mouse (total entry number, 6649). Approximately 80% of these proteins were categorized to the structural proteins of eukaryotic cells (cytoskeleton, nuclear matrix, cell adhesion, extracellular matrix, etc.), proteins involved in cellular signal transduction (receptor, channel, transporter, secondary signal transducing proteins, etc.), and nucleic acids-managing proteins (DNA polymerase, helicase, recombinase, transcription factor, etc.). These proteins are apparently related to cell shape determination, cell-cell communication, and modulation of cellular activities according to external signals. Thus, the data are consistent with the view that most of the large proteins are involved in characteristic functions of eukaryotic cells.

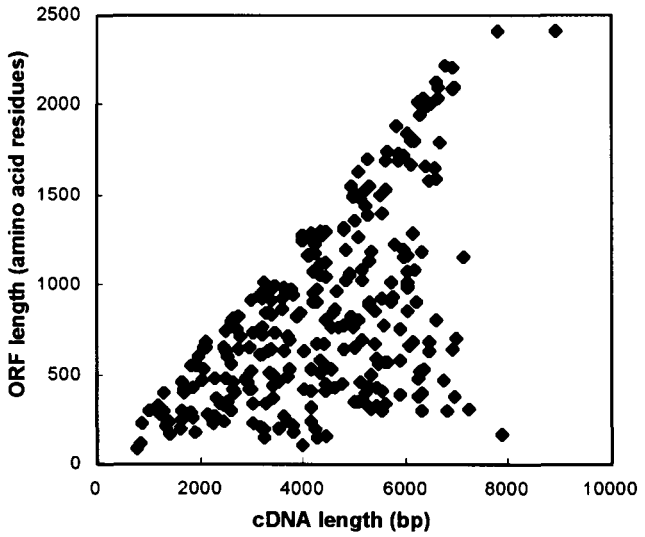

Figure 2. Relationship between the sizes of cDNAs and encoded proteins.
The sizes of deduced protein-coding regions for 280 human cDNA clones (KIAA0001 to KIAA0280[5]) were plotted against their cDNA sizes.

3.2. Correlation between cDNA size and its coding capacity

According to the sequence data obtained by our cDNA project,[5] there is a positive but nonlinear correlation between the length of cDNA and its coding capacity (Fig. 2). The coding capacity of cDNA increases with increasing cDNA size; the occurrence frequencies of cDNA clones encoding proteins larger than 1,000 amino acid residues are 5.8%, 35.7%, 45.7%, and 57.7% for cDNA sizes of 3–4 kb, 4–5 kb, 5–6 kb, and > 6 kb, respectively. Therefore, the minimum requirement for efficient analysis of cDNA clones encoding proteins longer than 1,000 amino acid residues is to construct a cDNA library enriched with clones longer than 4-kb insert sizes. The occurrence of such large cDNA clones in a library constructed by the conventional method is quite low, as described below. To circumvent this problem, we constructed a set of strictly size-fractionated cDNA libraries.

3.3. Construction and characterization of size-fractionated cDNA libraries

By repetition of the fractionation step of cDNA plasmids in a covalently closed circular form, we could prepare a set of human brain cDNA libraries (libraries 1 to 9), each being enriched with cDNA clones whose insert sizes are in a narrow range around their average sizes (Fig. 3A, columns 1 to 9). The integrity of cDNA clones in these libraries was examined by hybridization analysis with cDNA probes derived from 3 different transcripts of known sizes. The results are shown in Fig. 3B. With the neurodap1 (mRNA size: 5 kb) and α-spectrin (mRNA size: 8 kb) probes, the major bands were respectively detected in library 4 (average size: 5.3 kb) and library 7 (average size: 8.3 kb) as expected, although the libraries of smaller size were all shown to contain clones derived from these transcripts as truncated products. With the TPRD cDNA probe, which gives doublet bands of 8 and 9 kb by RNA blotting analysis, bands of expected sizes were hardly detected in libraries 8 and 9, and instead, bands obviously caused by rearrangement were observed. The persistent length of the reverse transcription with SuperScript II has been estimated to be around 7 kb.[16] Consistent with such estimation, the data in Fig. 3B suggest that the probability of containing artificial products increases in libraries larger than 8 kb. Separately, the occurrence of chimeric clones was also estimated by chromosomal mapping of 30 randomly sampled clones in library 5 with human/rodent hybrid cell panels: If a clone is mapped at the identical position using the sequence information on the 3′ and 5′ ends, the probability of being chimeric should be very low. The result of analysis showed that 29 out of 30 clones were mapped at the same loci (data not shown). Thus, we used libraries 4 to 6 for analysis of large proteins as those enriched with clones corresponding to full-length or nearly full-length transcripts, even though these libraries contain truncated products derived from larger transcripts as well.

Both the 3′- and 5′-terminal sequences of cDNA clones in library 6 were compared with those of a conventional cDNA library which has been prepared by the same procedure as used for the library construction described here except for size-fractionation of cDNAs (SuperScript™ human brain cDNA library, Gibco BRL, USA: average insert size, 1.38 kb). The sequences of randomly sampled cDNA clones were subjected to homology search using the BLAST software, and the highest scores for respective sequences were compiled. In this analysis, the sequences with the scores higher than 1,000 are categorized to the registered sequences in databases. When searched against the GenBank database without ESTs (release no. 96.0), the occurrence of the registered sequences in the conventional library was four to five times higher than those in library 6 for both the 3′- and 5′-termini (Fig. 4A, B). By comparison with the EST database compiled from the GenBank database, essentially similar patterns were obtained for the 5′-termini and the occurrence of the registered sequences in library 6 was very low compared with that in the conventional library (Fig. 4D). However, the contents of the registered sequences in the 3′-terminal sequences were around 40% to 50% for both libraries, indicating that the probability of "hitting" the registered ESTs at the 3′-terminal moiety is almost identical (Fig. 4C). The result clearly shows that the clones carrying unidentified sequences in the 5′-moiety have significantly been enriched in library 6 and the occurrence

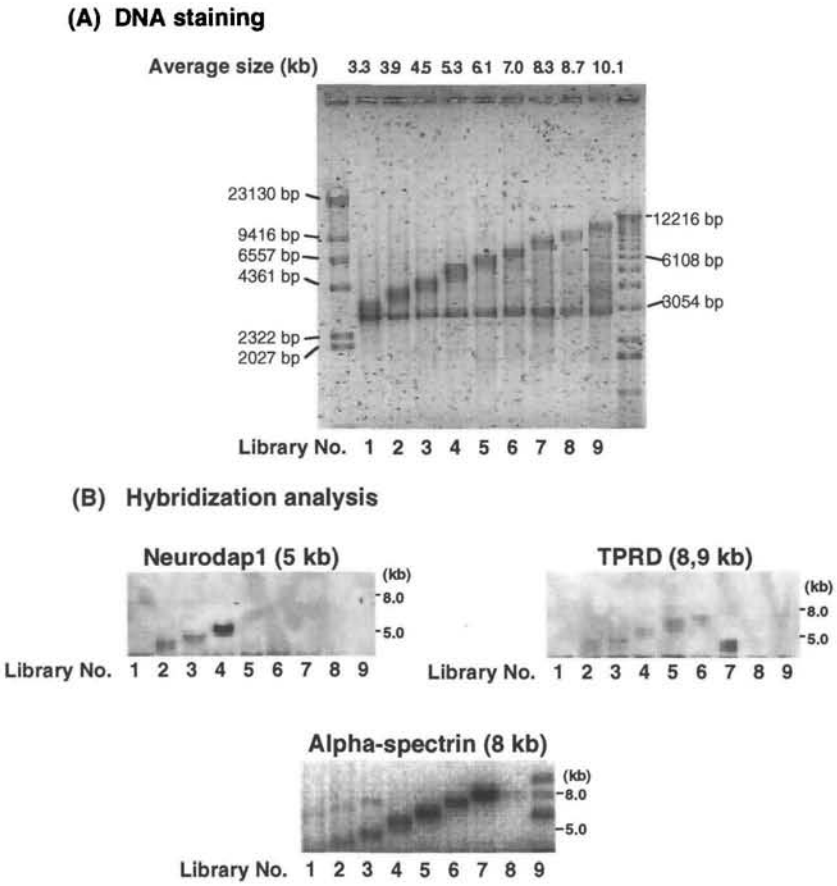

Figure 3. Hybridization analysis of the size-fractionated cDNA libraries with the probes for messenger RNAs of known sizes. In panel (A), plasmids recovered from respective libraries were digested with *Not* I and *Mlu* I, resolved on 0.7% agarose gel, and visualized by fluorescent staining with SYBR-Green I (Molecular Probe Inc., USA). The average insert sizes are given above the lanes. The 3.0-kb band commonly seen in each lane corresponds to pBluescript vector. Panel (B) displays the hybridization patterns of the *Not* I-*Mul* I digested cDNA plasmids from respective libraries with neurodap1, TPRD, and α-spectrin probes.

of such new clones is estimated to be more than 90%. Although we only examined the terminal sequences in library 6, our fractionated libraries, at least up to library 6, should greatly facilitate the characterization of unidentified large proteins. Although more than 600,000 ESTs have been deposited in databases and the number may increase more in the future, those EST data would be rather useful for sorting cDNAs with respect to their 3′-terminal moiety than for characterizing unidentified cDNA clones encoding large proteins.

3.4. Selection of cDNA clones coding for large proteins

To further validate the coding potentiality of cDNA clones in the size-fractionated libraries constructed, we adopted an *in vitro* transcription/translation system. Since the vector used for the construction of libraries bears a promoter for T7 RNA polymerase upstream from the 5′-end of the cDNA insert, the cDNA plasmids were directly added to the TNT T7-coupled reticulocyte lysate system and the products were analyzed on SDS-polyacrylamide gel. The result of this assay for 18 randomly sampled clones in library 6 are shown in Fig. 5 (lanes 1 to 18). Three out of 18 clones directed the synthesis of proteins larger than 100 kDa (lanes 2, 3, and 9), and 6 clones gave bands with apparent molecular masses of 50–100 kDa (lanes 4–6, 7, 14, and 16). The remaining clones did not produce any bands or proteins smaller than 50 kDa. Weak bands of around 70 kDa and 100 kDa, which were commonly seen in all the lanes, are intrinsic background products of this *in vitro* system. In lanes 20 and 21 in Fig. 5, the products generated from two cDNA clones, of which the potential coding regions have been predicted to be 1,382 and 982 amino acid residues from the sequence data, are shown (KIAA0139 and KIAA0144[17]). We have also determined the entire sequence of the cDNA clone of lane 3 and predicted its ORF size to be 1,451 amino acid residues (data not shown). Although multiple bands with differ-

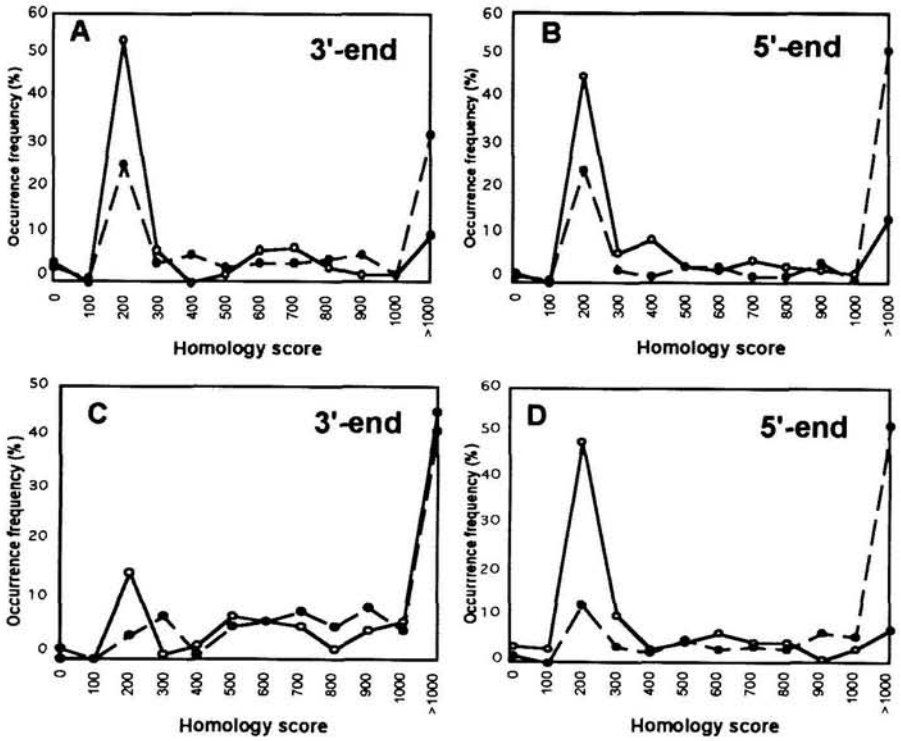

Figure 4. Homology search of the sequences at both the ends of cDNA clones in the size-fractionated and conventional cDNA libraries. The 3′-and 5′-end sequences of cDNA clones randomly selected (134 and 114 clones, respectively) from library 6 and the conventional cDNA library were subjected to homology search against the GenBank DNA database without ESTs or against the EST database generated from the GenBank database. The results of homology search of the 3′- and 5′-end sequences against the GenBank database are given in A and B, and those against the EST database in C and D, respectively. Solid lines: clones from library 6. Broken lines: clones from the conventional cDNA library.

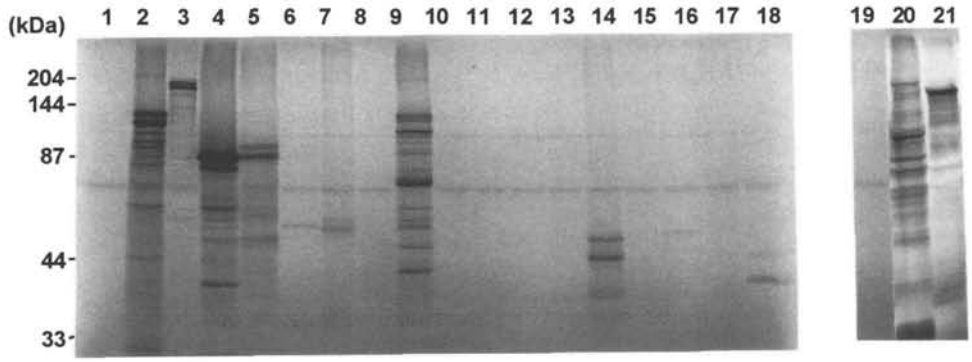

Figure 5. The protein products generated from 18 randomly sampled cDNA clones in the *in vitro* transcription/translation system. The SDS-polyacrylamide gel electrophoretic patterns of *in vitro* products generated from 18 cDNA clones in library 6 are shown in lanes 1 to 18, and those from two cDNA clones of known sequences (KIAA0139 and KIAA014117) in lanes 20 and 21. Lane 19 contains no DNA (control). The size markers are given on the left of the lanes.

ent signal strengths are seen in the corresponding lanes, the largest apparent molecular masses appeared to be in good agreement with the size of predicted ORFs within measurement error. So far, we have analyzed 100 cDNA clones and detected significant bands larger than 50 kDa in 20 clones (20%).

If this screening step is taken, the cDNA clones actually coding for large proteins can be experimentally selected. However, it is possible that a clone carrying an inert initiation signal in the *in vitro* system was overlooked. Alternatively, the possibility of detecting the translational initiation from internal ATG codons is also present.

As recently pointed out by Kozak,[18] identification of the authentic translational initiation sites in cDNA must be done very carefully, otherwise mis-assignment will result. In this respect, the data on the *in vitro* protein products should offer valuable information for the initiation site assignment.

Using the size-fractionated libraries constructed as described above (libraries 4, 5, and 6), screening of new cDNA clones encoding large proteins and chromosomal mapping as well as entire sequence analysis of the screened clones are in progress.

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, and Kazuko Yamada, and Naoko Suzuki for their technical assistance.

References

1. Adams, M. D., Kerlavage, A. R., Fleischmann, R. D. et al. 1995, Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence, *Nature*, **377** (supplement), 3–17.
2. Houlgatte, R., Mariage, S. R., Duprat, S. et al. 1995, The Genexpress Index: a resource for gene discovery and the genic map of the human genome, *Genome Res*, **5**, 272–304.
3. Hillier, L., Lennon, G., Becker, M. et al. 1996, Generation and analysis of 280,000 human expressed sequence tags, *Genome Res.*, **6**, 807–828.
4. Nagase, T., Seki, N., Ishikawa, K.-I., Tanaka, A., and Nomura, N. 1996, Prediction of the coding sequences of unidentified human genes. V. The coding sequences of 40 new genes (KIAA0161-0200) deduced by analysis of cDNA clones from human cell line KG-1, *DNA Res.*, **3**, 17–24.
5. 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.
6. Nakayama, M., Miyake, T., Gahara, Y., Ohara, O., and Kitamura, T. 1995, A novel RING-H2 motif protein downregulated by axotomy: Its characteristic localization at the postsynaptic density of axosomatic synapse, *J. Neurosci.*, **15**, 5238–5248.
7. Ohira, M., Ootsuyama, A., Suzuki, E. et al. 1996, Identification of a novel human gene containing the tetratricopeptide repeat domain from the Down syndrome region of chromosome 21, *DNA Res.*, **3**, 9–16.
8. Moon, R. T. and McMahon, A. P. 1990, Generation of diversity in nonerythroid spectrins. Multiple polypeptides are predicted by sequence analysis of cDNAs encompassing the coding region of human nonerythroid alpha-spectrin, *J. Biol. Chem.*, **265**, 4427–4433.
9. Church, G. M. and Gilbert, W. 1984, Genomic sequencing, *Proc. Natl. Acad. Sci. USA*, **81**, 1991–1995.
10. Devereux, J., Haeberli, P., and Smithies, O. 1984, A comprehensive set of sequence analysis programs for the VAX, *Nucleic Acids Res.*, **12**, 387–395.
11. Laemmli, U. K. 1970, Cleavage of structural proteins during the assembly of the head of bacteriophage T4, *Nature*, **277**, 680–685.
12. Fleischmann, R. D., Adams, M. D., White, O. et al. 1995, Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd, *Science*, **269**, 496–512.
13. Fraser, C. M., Gocayne, J. D., White, O. et al. 1995, The minimal gene complement of *Mycoplasma genitalium*, *Science*, **270**, 397–403.
14. Kaneko, T., Sato, S., Kotani, H. et al. 1996, Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. Strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding region, *DNA Res.*, **3**, 109–136.
15. Bult, C. J., White, O., Olsen, G. J. et al. 1996, Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*, *Science*, **273**, 1058–1073.
16. Gerard, G. F., Schmidt, B. J., Kotewicz, M. L., and Campbell, J. H. 1992, cDNA synthesis by moloney murine leukemia virus RNase H-minus reverse transcriptase possessing full DNA polymerase activity, *Focus*, **14**, 91–93.
17. Nagase, T., Seki, N., Tanaka, A., Ishikawa, K., and Nomura, N. 1995, Prediction of the coding sequences of unidentified human genes. IV. The coding sequences of 40 new genes (KIAA0121-KIAA0160) deduced by analysis of cDNA clones from human cell line KG-1, *DNA Res*, **2**, 167–174.
18. Kozak, M. 1996, Interpreting cDNA sequences: some insights from studies on translation, *Mammalian Genome*, **7**, 563–574.