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

Takahiro Nagase,* Reiko Kikuno, Ken-ichi Ishikawa, Makoto Hirozawa, and Osamu Ohara

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

(Received 31 March 2000)

Abstract

To provide information regarding the coding sequences of unidentified human genes, we have conducted a sequencing project of human cDNAs which encode large proteins. We herein present the entire sequences of 100 cDNA clones of unknown human genes, named KIAA1444 to KIAA1543, from two sets of size-fractionated human adult and fetal brain cDNA libraries. The average sizes of the inserts and corresponding open reading frames of cDNA clones analyzed here were 4.4 kb and 2.6 kb (856 amino acid residues), respectively. Database searches of the predicted amino acid sequences classified 53 predicted gene products into the following five functional categories: cell signaling/communication, nucleic acid management, cell structure/motility, protein management and metabolism. It was also revealed that homologues for 32 KIAA gene products were detected in the databases, which were similar in sequence through almost their entire regions. Additionally, the chromosomal loci of the genes were determined by using human-rodent hybrid panels unless their chromosomal loci were already assigned in the public databases. The expression levels of the genes were monitored in spinal cord, fetal brain and fetal liver, as well as in 10 human tissues and 8 brain regions, by reverse transcription-coupled polymerase chain reaction, products of which were quantified by enzyme-linked immunosorbent assay.

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

In December of 1999, the complete DNA sequence covering almost the entire region of the euchromatic part of human chromosome 22 was reported. Moreover, in the near future, a working draft sequence, covering at least 90% of the human genome, is expected to become available as a result of international collaboration of human genome sequencing. However, it is difficult to conclusively identify genes only from the genomic sequence by any current computer programs because of complicated and extensive splicing events during transcription and our limited knowledge of the sequence signals which define the transcriptional start and termination sites. Therefore, it is still important to collect full-length cDNAs and analyze their sequences, not only for utilization for functional studies but also for interpretation of the gene structure, even after the human genome sequence becomes available. Considering this situation, we have been making efforts to accumulate information on the coding sequences of unidentified human genes. Currently, we have focused our sequencing efforts on the unidentified genes encoding large proteins in human brain since these gene products appear to play important roles in the central nervous system. As an extension of the preceding studies, we herein report the predicted coding sequence of 100 new cDNA clones which have the potential to code for large proteins in vitro. In addition to the specific features of the newly predicted protein sequences annotated by the database search, the expression profiles and the chromosomal locations of these 100 new genes are also reported. The growing catalog of genes encoding large proteins should provide a wealth of information concerning the primary structures of proteins present in the human brain that are difficult to identify by conventional methods of gene discovery.

1. Sequence Analysis and Prediction of Protein-Coding Regions in cDNA Clones

The cDNA clones were isolated from the size-fractionated human adult brain cDNA libraries Nos. 2 to 5 (insert sizes ranging from 4 to 6 kb) and
Figure 1. Physical maps of cDNA clones analyzed. The physical maps shown here were constructed from the sequence data of respective cDNA clones or, when necessary, from the combination of cDNA clones and RT-PCR products. 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, with or without the contexts of 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 | cDNA length (bp) | ORF length (amino acid residues) | Chromosomal location |
|--------------------|------------------|------------------|----------------------------------|---------------------|
| 1444               | AB040877         | 1248             | 416                               | X                   |
| 1445               | AB040878         | 559              | 1202                              | 3                   |
| 1446               | AB040879         | 5806             | 645                               | 14                  |
| 1447               | AB040880         | 5166             | 1721                              | 17                  |
| 1448               | AB040881         | 5336             | 526                               | 1                   |
| 1449               | AB040882         | 5314             | 607                               | 3                   |
| 1450               | AB040883         | 5750             | 1139                              | 4                   |
| 1451               | AB040884         | 5556             | 488                               | 12                  |
| 1452               | AB040885         | 8800             | 694                               | 16                  |
| 1453               | AB040886         | 5879             | 1123                              | 17                  |
| 1454               | AB040887         | 5438             | 1265                              | 15                  |
| 1455               | AB040888         | 5309             | 1769                              | 4                   |
| 1456               | AB040889         | 7595             | 421                               | 8                   |
| 1457               | AB040890         | 5517             | 989                               | 12                  |
| 1458               | AB040891         | 5843             | 612                               | 4                   |
| 1459               | AB040892         | 6417             | 853                               | 1                   |
| 1460               | AB040893         | 5273             | 1400                              | 16                  |
| 1461               | AB040894         | 5285             | 1498                              | 2                   |
| 1462               | AB040895         | 5104             | 800                               | 10                  |
| 1463               | AB040896         | 5437             | 532                               | 12                  |
| 1464               | AB040897         | 5224             | 621                               | 16                  |
| 1465               | AB040898         | 5737             | 642                               | 15                  |
| 1466               | AB040899         | 5520             | 505                               | 7                   |
| 1467               | AB040900         | 6124             | 432                               | 12                  |
| 1468               | AB040901         | 4661             | 985                               | 18                  |
| 1469               | AB040902         | 4723             | 662                               | 20                  |
| 1470               | AB040903         | 4028             | 564                               | 16                  |
| 1471               | AB040904         | 4038             | 1345                              | 3                   |
| 1472               | AB040905         | 5258             | 605                               | 8                   |
| 1473               | AB040906         | 4133             | 574                               | 19                  |
| 1474               | AB040907         | 1692             | 1079                              | 19                  |
| 1475               | AB040908         | 3869             | 986                               | 15                  |
| 1476               | AB040909         | 4878             | 1220                              | 2                   |
| 1477               | AB040910         | 4638             | 870                               | 1                   |
| 1478               | AB040911         | 5818             | 570                               | 12                  |
| 1479               | AB040912         | 4746             | 464                               | 15                  |
| 1480               | AB040913         | 3154             | 682                               | X                   |
| 1481               | AB040914         | 4755             | 1380                              | 4                   |
| 1482               | AB040915         | 4841             | 813                               | 4                   |
| 1483               | AB040916         | 2692             | 428                               | 6                   |
| 1484               | AB040917         | 3174             | 700                               | 19                  |
| 1485               | AB040918         | 4006             | 1104                              | 8                  |
| 1486               | AB040919         | 4498             | 677                               | 2                   |
| 1487               | AB040920         | 3992             | 650                               | 4                   |
| 1488               | AB040921         | 3060             | 852                               | 3                   |
| 1489               | AB040922         | 4330             | 511                               | 7                   |
| 1490               | AB040923         | 4115             | 749                               | 13                  |
| 1491               | AB040924         | 2920             | 757                               | 9                   |
| 1492               | AB040925         | 4323             | 711                               | 2                   |
| 1493               | AB040926         | 4768             | 415                               | 11                  |

a) Accession numbers of DDBJ, EMBL and GenBank databases. b) Values excluding poly(A) sequences. c) Chromosome numbers identified by using GeneBridge 4 radiation hybrid panel unless specified. 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/hug.

The chromosomal locations highlighted by asterisks were fetched from the UniGene database. The chromosomal locations highlighted by sharp were referred from the GeneBank database because the sequences of the cDNA clones could be found in the genomic sequences whose chromosome numbers were assigned. The actual primer sequences and the PCR conditions used for the radiation hybrid mapping are also accessible through the World Wide Web site referred above. d) Chromosome number determined by using CCR human-rodent hybrid panel. e) cDNA and ORF lengths were revised by direct analysis of the RT-PCR products. f) Nucleotide sequences were determined after subcloning of the internal Not I-digested fragment. Therefore, cDNA length of these genes represented those of internal Not I-digested fragment. g) cDNA clones were selected by analysis of 5'-end single-pass sequences by the computer-assisted method.
performed according to the methods previously described in detail. Thirty clones (KIAA1509-KIAA1538) seemed to carry spurious coding interruption caused by errors of the reverse transcriptase or by retained intron sequences. For these cases, the sequences of the regions causing interruption of open reading frame (ORF) were reexamined by direct sequencing of the major products of reverse transcription-coupled polymerase chain reaction (RT-PCR) to predict authentic protein-coding sequences in brain. As the results of these confirmations, spurious interruptions were found in the following cDNA clones: ORFs in 23 clones (KIAA1509-KIAA1512, KIAA1514-KIAA1517, KIAA1519-KIAA1521, KIAA1523, KIAA1524, KIAA1526-KIAA1528, KIAA1530 and KIAA1532-KIAA1537) were found to carry single or multiple insertions, most of which probably corresponded to retained intronic sequences; ORFs in 10 clones (KIAA1513, KIAA1515, KIAA1520-KIAA1522, KIAA1524, KIAA1525, KIAA1529, KIAA1531 and KIAA1538) were found to carry single or multiple deletions; ORFs in 2 clones (KIAA1523 and KIAA1524) were frame-shifted by a single nucleotide deletion. For those genes, the revised sequences by the RT-PCR experiments, not the actual cloned cDNA sequences, were deposited to GenBank/EMBL/DDBJ.

Table 2. Functional classifications of the gene products.

| Gene product | GO: ID | GO: ID | Gene product | GO: ID |
|--------------|-------|-------|--------------|-------|
| 143          |       |       | 145          |       |

The size-fractionated human fetal brain cDNA libraries Nos. 4 and 6 (insert sizes ranging from 4 to 7 kb) previously constructed. In this report, 20 cDNA clones (KIA1497-KIA1508, KIA1528-KIA1531, KIA1538 and KIAA1541-KIA1543) were selected from the adult brain libraries and the remaining 20 cDNA clones were obtained from the fetal brain cDNA libraries. According to our selection system for unidentified genes, the clones with unidentified sequences at both ends were chosen by single-pass sequencing and homology search against the GenBank database (release 116.0) excluding expressed sequence tags and genomic sequences. For selection of cDNA clones to be entirely sequenced, we used an in vitro transcription/translation system to examine their protein-coding capacities and/or a computer-based method based on GeneMark analysis to predict protein-coding potentialities on their 5’-end sequences in this study. Eighty-eight cDNA clones were selected by the in vitro expression system, while 12 cDNA clones (KIAA1532-KIAA1543) were chosen by the computer-assisted method. Entire sequencing of these clones was performed according to the methods previously described in detail. Thirty clones (KIAA1519-KIAA1538) seemed...
2-2. Predicted function by motif search

| Position | Gene product | Pfam ID | E-value | Definition |
|-----------|--------------|---------|---------|------------|
| Cell signaling/communication | KIAA1405 | PP01467 | 1.00E-14 | Leucine rich repeat C-terminal domain |
| | KIAA1406 | PP01469 | 3.90E-12 | Leucine Rich Repeat |
| | KIAA1407 | PP01463 | 1.00E-09 | Immunoglobulin domain |
| | KIAA1408 | PP01463 | 3.00E-09 | Leucine Rich Repeat |
| | KIAA1409 | PP01463 | 8.00E-10 | Leucine Rich Repeat |
| | KIAA1410 | PP01463 | 2.40E-06 | Filamin family |
| | | PP01463 | 4.70E-06 | Filamin family |
| | | PP01463 | 8.50E-05 | Leucine rich repeat C-terminal domain |
| | KIAA1411 | PP01463 | 4.00E-10 | SH3 domain |
| | KIAA1412 | PP01463 | 1.00E-11 | SH3 domain |
| | KIAA1413 | PP01463 | 6.70E-06 | Leucine Rich Repeat |
| | | PP01463 | 1.00E-09 | Leucine Rich Repeat |
| | | PP01463 | 3.90E-09 | Leucine Rich Repeat |
| | | PP01463 | 9.60E-11 | Leucine Rich Repeat |
| | | PP01463 | 1.50E-09 | Immunoglobulin domain |
| | | PP01463 | 6.70E-06 | Leucine rich repeat N-terminal domain |
| | | PP01463 | 8.40E-03 | Immunoglobulin domain |
| | KIAA1414 | PP01463 | 1.30E-12 | Leucine rich repeat |
| | | PP01463 | 1.30E-07 | Filamin |
| | | PP01463 | 4.20E-15 | Filamin |
| | | PP01463 | 3.90E-11 | Filamin |
| | | PP01463 | 3.70E-16 | Filamin |
| | | PP01463 | 6.50E-10 | Filamin |
| | | PP01463 | 8.40E-14 | Filamin |
| | | PP01463 | 5.60E-15 | Filamin |
| | | PP01463 | 1.30E-12 | Filamin |
| | | PP01463 | 2.30E-06 | Filamin |
| | | PP01463 | 7.00E-14 | Filamin |
| | | PP01463 | 3.00E-04 | Filamin |
| | | PP01463 | 5.90E-10 | Filamin |
| | | PP01463 | 5.40E-07 | Filamin |
| | KIAA1415 | PP01463 | 1.00E-09 | Leucine rich repeat |
| | | PP01463 | 4.00E-03 | Leucine rich repeat |
| | | PP01463 | 7.00E-01 | Immunoglobulin domain |
| | | PP01463 | 1.70E-03 | 2 transmembrane receptor (Secretin family) |
| Nucleic acid management | KIAA1416 | PP00076 | 1.70E-02 | RNA recognition motif |
| | KIAA1417 | PP00076 | 4.60E-02 | Zinc finger, C2H2 type |
| | | PP00076 | 9.40E-02 | Zinc finger, C2H2 type |
| | | PP00076 | 3.80E-03 | Zinc finger, C2H2 type |
| | | PP00076 | 1.00E-17 | Bromodomains |
| | KIAA1418 | PP00439 | 3.00E-13 | Bromodomains |
| | | PP00439 | 8.90E-18 | PRD finger |
| | KIAA1419 | PP00439 | 2.70E-01 | Rho family |
| | | PP00439 | 4.90E-01 | Rho family |
| | | PP00439 | 5.90E-01 | Zinc finger, C2H2 type |
| | KIAA1420 | PP00439 | 1.50E-19 | Zn finger, C2H2 type |
| Protein management | KIAA1421 | PP00439 | 3.70E-06 | Leucine rich repeat C-terminal domain |
| | | PP00439 | 8.00E-01 | TPR domain |

a) Motif search was performed by HMMER2.1.1 against Pfam database (release 5.1). b) Function was classified based on the annotation of the Pfam entry which was hit in the query sequence. c) Only the entries possessing the expectation value (E-value) less than 1.0 were presented.

databases and used for prediction of their protein-coding sequences unless otherwise stated. The results of the comparison between the cloned DNA and the revised DNA sequences are available through the World Wide Web site at http://www.kazusa.or.jp/huge. Notably, clones for seven genes (KIAA1444, KIAA1447, KIAA1453, KIAA1471, KIAA1498, KIAA1502 and KIAA1506) seemed to lack regions encoding C-terminal portions due to the presence of a Not I site in their coding regions because cDNAs were digested with Not I before ligation to a vector. In contrast, clones for two genes (KIAA1445 andKIAA1450) were found to lack the 5'-portions of the sequences due to the presence of an internal Not I site. For these genes, the nucleotide sequences of only the region between two Not I sites were determined, since their original clones were most likely to harbor intermolecularly ligated two independent cDNAs. After these revisions, the average size of the cDNA sequences reached 4.4 kb and that of the predicted coding region was approximately 856 amino acid residues. Physical maps of the 100 cDNA sequences 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. Repeat sequences are also shown in Fig. 1. Eleven genes had 5'-untranslated regions longer than 1 kb. We could not completely rule out the possibility that these clones retained a 5'-intron upstream of the predicted protein coding region, since the RNAs used to construct the cDNA libraries contained heterogeneous nuclear RNAs besides the cytoplasmic mRNAs. Chromosomal loci of 54 newly identified genes were determined by using human-rodent hybrid panels, GeneBridge 4 (Research Genetics Inc., USA) or CCR (Coriell Cell Repositories, USA) since their mapping data were not available. The chromosomal locations of the 42 genes, which are highlighted by asterisks in Table 1, were fetched from the UniGene database (http://www.ncbi.nlm.nih.gov/UniGene). The chromosomal locations of the remaining four genes, which are highlighted by sharp (#) in Table 1, were taken from the description of the GenBank database.
Table 3. Homologies of the newly identified genes found in various databases.a)

| Database/ID of predicted protein sequences from yeast 10 and yeast-protein/yeast_nrpep.fasta.Z, ftp.sanger.ac.uk:/pub/databases/C.elegans_sequences/C_elegans_proteins_1998-10-16.pep | | |
|---|---|---|
| HUGO and new genes KIAA1455 | 3136 | KIAA1127 |
| KIAA1466 | 1406 | KIAA1193 |
| KIAA1469 | 862 | KIAA0405 |
| KIAA1473 | 574 | KIAA0661 |
| KIAA1474 | 1220 | KIAA0314 |
| KIAA1480 | 682 | KIAA0951 |
| KIAA1487 | 650 | KIAA1021 |
| KIAA1528 | 740 | KIAA0937 |
| KIAA1531 | 668 | KIAA1221 |

a) The definition of homologues used here was the proteins found in the databases satisfying the following conditions: i) the length ranged from 80% to 125% of the query sequence; ii) the ratio of the length of aligned region to that of the query sequence was the same to that explained in Table 2-1. b) The following databases were used. HUGE, our cDNA-encoded protein database (http://www.kazusa.or.jp/huge); yeast, non redundant peptide database from genome-ftp.stanford.edu: C. elegans, protein database deduced from C. elegans full genome sequence (ftp.sanger.ac.uk:/pub/databases/C.elegans_sequences/C.elegans.proteins_1998-10-16.pep) and the entries derived from C. elegans of OWL, and OWL (release 31.4). In the case of database search against the following public databases: pro-
Figure 2. 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 according to the methods previously described in detail. Gene names are given as KIAA numbers at the left side of each set of color codes. Tissue and brain region names are indicated above the top sets of color codes. A color conversion panel shown at the bottom 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. Besides 10 tissues, 9 regions of the adult central nervous system (amygdala, corpus callosum, cerebellum, caudate nucleus, hippocampus, substantia nigra, subthalamus, thalamus, and spinal cord) and fetal brain were included in the expression profiling. As a control, mRNA levels in fetal liver were also examined.
of functional motifs/domains, since they did not show sequence similarity to functionally annotated proteins (Table 2-2). In total, 48 gene products (91% of genes functionally annotated here) were suggested to have functions relating to cell signaling/communication, nucleic acid management or cell structure/motility. To find the genes structurally conserved in others, we tentatively defined “homologues” as genes with at least 30% amino acid identity spanning almost the entire region (more than 80% coverage against the query protein sequence). As shown in Table 3, 32 KIAA gene products were found to have “homologues” in the databases.

3. Expression Profiles of Predicted Genes

The expression profiles of the genes newly identified in this study are represented in Fig. 2 by color codes. The expression levels of six genes (KIAA1444, KIAA1446, KIAA1472, KIAA1479, KIAA1484 and KIAA1497) were relatively higher in brain than in other tissues. Among them, KIAA1446 and KIAA1497 were homologues of rat brain-enriched guanylate kinase-associated protein and mouse neuronal leucine-rich repeat protein 1, respectively. The gene product of KIAA1472 exhibited similarity in sequence through almost the entire region (more than 80% coverage against the query protein sequence). As defined “homologues” as genes with at least 30% amino acid identity similarity spanning almost the entire region (more than 80% coverage against the query protein sequence).

Acknowledgements: This project was supported by grants from the Kazusa DNA Research Institute. We thank Tomomi Tajima, Keishi Ozawa, Tomomi Kato, Kazuhiro Sato, Akiko Ukigai, Emiko Suzuki, Kazuko Yamada, Kiyoe Sumi, Takashi Watanabe, Kozejane, Naoko Shibano, Kazutaka Mitsui, Mina Waki and Nobue Kashima for their technical assistance.

References

1. Dunham, I., Shimizu, N., Roe, B. A. et al. 1999, The DNA sequence of human chromosome 22, Nature, 402, 489–495.
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–KIAA0040) deduced by analysis of randomly sampled cDNA clones from human immature myeloid cell line KG-1, DNA Res., 1, 27–35.
3. Nagase, T., Kikuno, R., Ishikawa, K.-I. et al. 2000, Prediction of the coding sequences of unidentified human genes. XVI. The complete sequences of 150 new cDNA clones from brain which code for large proteins in vitro, DNA Res., 7, 65–73.
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. Hirosawa, M., Nagase, T., Ishikawa, K.-I., Kikuno, R., Nomura, N., and Ohara, O. 1999, Characterization of cDNA clones selected by the GeneMark analysis from size-fractionated cDNA libraries from human brain, DNA Res., 6, 329–336.
6. 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.
7. Nagase, T., Ishikawa, K.-I., Miyajima, N. et al. 1998, Prediction of the coding sequences of unidentified human genes. IX. 100 new cDNA clones from brain which can code for large proteins in vitro, DNA Res., 5, 31–39.
8. Gyapay, G., Schmitt, K., Fizames, C. et al. 1996, A radiation hybrid map of the human genome, Hum. Mol. Genet., 5, 339–346.
9. Bleasby, A. J., Akrigg, D., and Attwood, T. K. 1994, OWL - a non-redundant composite protein sequence database, Nucleic Acids Res., 22, 3574–3577.
10. Goffeau, A., Barrell, B. G., Bussey, H. et al. 1996, Life with 6000 genes, Science, 274, 546–567.
11. The C. elegans Sequencing Consortium. 1998, Genome sequence of the nematode C. elegans: A platform for investigating biology, Science, 282, 2012–2018.
12. Bateman, A., Birney, E., Durbin, R. et al. 1999, Pias 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins, Nucleic Acids Res., 27, 260–262.
13. Kikuno, R., Nagase, T., Suyama, M., Waki, M., Hirosawa, M., and Ohara, O. 2000, HUGE: a database for human large proteins identified in Kazusa cDNA sequencing project, Nucleic Acids Res., 28, 331–332.
14. 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, 277–286.
15. Deguchi, M., Hata, Y., Takeuchi, M. et al. 1998, BE-GAIN (brain-enriched guanylate kinase-associated protein), a novel neuronal PSD-95/SAP90-binding protein, J. Biol. Chem., 273, 26269–26272.
16. Taguchi, A., Wanaka, A., Mori, T. et al. 1996, Molecular cloning of novel leucine-rich repeat proteins and their expression in the developing mouse nervous system, Brain Res. Mol. Brain Res., 35, 31–40.