338–339 *Nucleic Acids Research, 1999, Vol. 27, No. 1*

HUGE: a database for human large proteins identified by Kazusa cDNA sequencing project

Mikita Suyama, Takahiro Nagase and Osamu Ohara*

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

Received September 3, 1998; Revised September 24, 1998; Accepted October 13, 1998

ABSTRACT

HUGE is a database for human large proteins newly identified by Kazusa cDNA project, which aims to predict protein primary structures from sequences of human large cDNAs (>4 kb). In particular, cDNA clones capable of coding for large proteins (>50 kDa) are current targets of the project. More than 700 sequences of human cDNAs (average size, 5.1 kb) have been determined to date and deposited in the public databases. Notable information implied from the cDNAs and the predicted protein sequences can be obtained through HUGE via the World Wide Web at URL http://www.kazusa.or.jp/huge

INTRODUCTION

Kazusa DNA Research Institute has been conducting a cDNA sequencing project for prediction of primary structures of unidentified human proteins. In particular, we have been interested in long cDNA clones which direct synthesis of large proteins (>50 kDa) (1). To date, we have deposited more than 700 human cDNA sequences (average size: 5.1 kb) in public databases (2). Substantial increment of our cDNA sequence data has prompted us to generate a database for protein sequences predicted by the cDNA analysis, as future functional studies using these cDNAs would inevitably require systematic and comprehensive overview of the predicted protein sequence data. In this context, this database, called 'HUGE' (Human Unidentified Gene-Encoded large protein database), was constructed to provide more detailed information concerning the predicted primary structures by the Kazusa cDNA project than those retrievable from the public databases. Since we make cDNA clones publicly available for research purposes, once the sequence data are deposited to the GenBank/EMBL/DDBJ databases, HUGE is also expected to provide practically important information of clones to worldwide clone users. While HUGE focuses mainly on the characteristics of predicted primary structures, other important information concerning cDNA clones from the genomic viewpoint is compiled in the Kazusa human cDNA database at http://www.kazusa.or.jp/cDNA

CONTENTS OF GENE/PROTEIN CHARACTERISTIC TABLE

Because all genes newly characterized by the Kazusa cDNA project are conventionally identified by KIAA plus a four figure number, these KIAA numbers are used as primary gene identifiers in HUGE. Gene/protein characteristic tables summarize the results of sequence analyses of cloned cDNAs and the predicted proteins. Each HUGE entry has its own table (Fig. 1). There are currently more than 700 gene/protein characteristic tables in HUGE.

The table begins with a section indicating the accession number of the cDNA sequence in the public database, the alias name of the gene, the clone name, and the biological source of the cDNA library from which the clone was isolated.

The next section describes characteristics of the cloned cDNA sequence and includes four subsections. The first two subsections show the length of the cloned DNA sequence and the physical map constructed from the actual sequence data of the isolated cDNA clone. The open reading frame and untranslated regions are shown by solid and open boxes, respectively, and the position of the first ATG codon is indicated by a triangle. Alu and other repetitive sequences detected by RepeatMasker (http://ftp.genome.washington.edu/RM/RepeatMasker.html) are also displayed by dotted and hatched boxes, respectively, in the physical map. The next subsection offers the restriction map of the isolated cDNA generated by using the list of commercially available restriction enzymes (3). The last subsection describes prediction of the protein coding region by GeneMark analysis and linked to the graphical output of the GeneMark-RC analysis (4). The GeneMark analysis also gives warnings for N-terminal truncation and for spurious interruption of the coding region, if detected. As for the clones warned by the GeneMark analysis, we performed additional experiments using reverse transcription-coupled polymerase chain reaction (RT-PCR) to eliminate artifacts in cloning. Details of the evaluation of the cDNA sequences by the GeneMark analysis will be reported elsewhere (manuscript in preparation).

The third section, in which we overview various characteristics of the predicted protein sequences, is divided into five subsections. The length of the predicted protein sequence is indicated in the first subsection. The second subsection, which is optional, describes whether the translation is conceptual or not; when error(s) in cloning (e.g., frameshift, or nonsense mutation, or retention of intron) were experimentally detected in the clone actually sequenced, translation was carried out not for the sequence of the isolated cDNA but for the experimentally corrected one. The next two subsections show the results of the homology searches against OWL database (5) and other HUGE entries. The top five entries given expectation values less than 0.001 by FASTA (6) are aligned along the query sequence in a graphical overview. Numbers at the left and the right sides of black lines in the overview indicate the numbers of amino acid residues of non-homologous N-terminal and C-terminal portions of the homologous entries, respectively. The FASTA output and the multiple alignment of these entries can

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

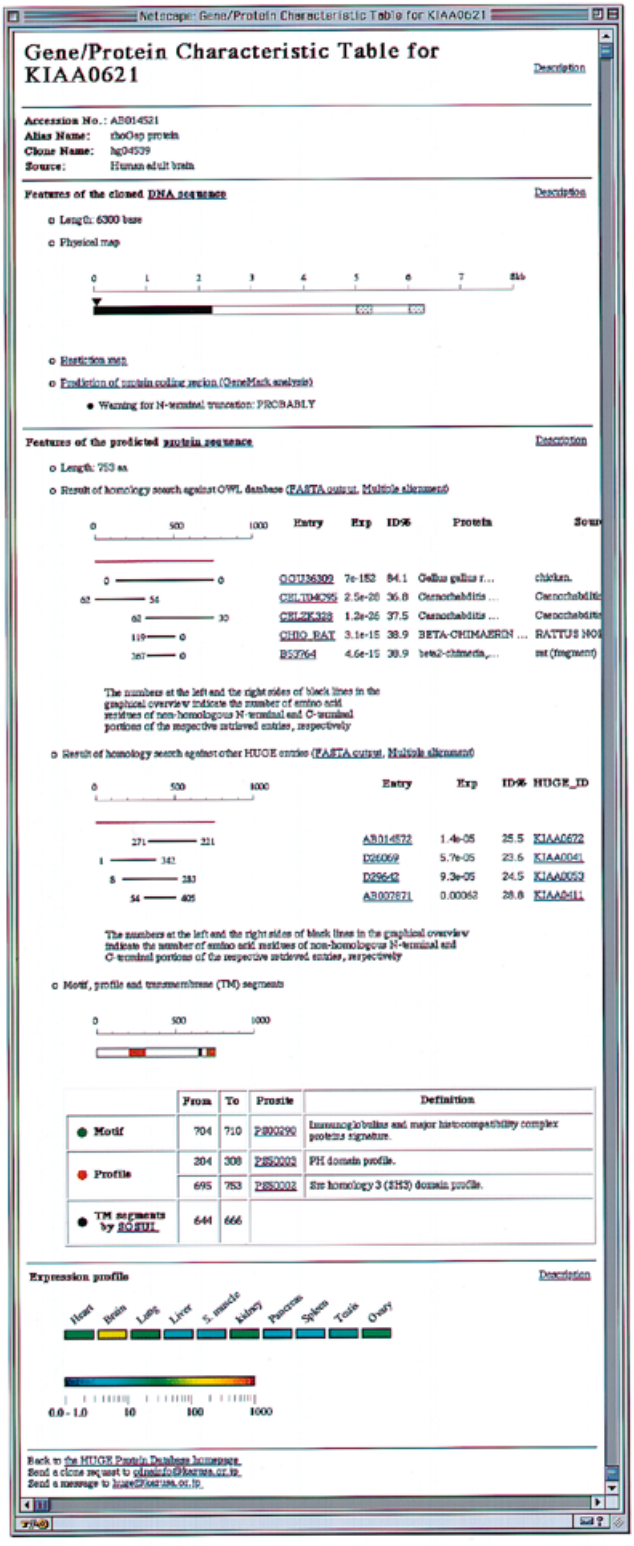

Figure 1. A typical overview of a gene/protein characteristic table in HUGE. This gene/protein characteristic table is for KIAA0621. There are some links to raw data of the sequence analyses (e.g., the GeneMark coding prediction and FASTA homology search) done for this cDNA and protein. See text for a description of the fields.

be viewed by clicking. The last subsection illustrates the results of the motif/profile analysis and the prediction of transmembrane helical segments. Although the PROSITE database (7) was used for the motif analysis, the following relaxed sequence motifs were excluded from the analysis because they appear too often and are considered to be less informative: amidation site, N-glycosylation site, cAMP- and cGMP-dependent protein kinase phosphorylation site, casein kinase II phosphorylation site, protein kinase C phosphorylation site and tyrosine kinase phosphorylation site. The profile analysis was conducted with profile entries in the PROSITE database by using the pfscan program in the pftools package (ftp://ulrec3.unil.ch/pub/pftools). Membrane-spanning regions were predicted by the SOSUI program (8).

The last section is optional and shows the expression pattern at the mRNA level determined by RT-PCR coupled with enzyme-linked immunosorbent assay (ELISA). By using external control reactions with the authentic plasmid, the mRNA levels are expressed as equivalent amounts of the authentic plasmid DNA (fg) per ng of poly(A)$^+$ RNA. For an at-a-glance overview, the mRNA levels are displayed in colors using the digit-color conversion panel shown in this section.

HOW TO ACCESS GENE/PROTEIN CHARACTERISTIC TABLES OF INTEREST

The home page allows users to easily reach a gene/protein characteristic table of interest by three different approaches. The first is to directly enter the list of gene/protein characteristic tables. The second is to search for tables that contain query keywords. Tables thus found can be further confined by adding another keyword one by one. The search can be carried out not only on the entire fields in the gene/protein characteristic table but also on a specified field such as motif/profile information and FASTA results. It is also possible to retrieve HUGE entries according to the size of cDNAs/proteins and the number of the predicted transmembrane segments. As the third approach, it is also possible to search for a gene/protein characteristic table of interest by the FASTA homology search from a user's query sequence (either nucleotide or amino acid sequence) against HUGE.

ACKNOWLEDGEMENTS

We thank Takatsugu Hirokawa, Seah Boon-Chieng, and Shigeki Mitaku for allowing us to use the SOSUI program for prediction of transmembrane helical regions. We also thank Makoto Hirosawa for providing us with the results of GeneMark-RC analysis. This work was supported by the Kazusa DNA Research Institute Foundation.

REFERENCES

1 Ohara,O., Nagase,T., Ishikawa,K.-I., Nakajima,D., Ohira,M., Seki,N. and Nomura,N. (1997) *DNA Res.*, **4**, 53–59.
2 Ishikawa,K.-I., Nagase,T., Suyama,M., Miyajima,N., Tanaka,A., Kotani,H., Nomura,N. and Ohara,O. (1998) *DNA Res.*, **5**, 169–176.
3 Roberts,R.J. and Macelis,D. (1998) *Nucleic Acids Res.*, **26**, 338–350.
4 Hirosawa,M., Isono,K., Hayes,W. and Borodovsky,M. (1997) *DNA Seq.*, **8**, 17–29.
5 Bleasby,A.J., Akrigg,D. and Attwood,T.K. (1994) *Nucleic Acids Res.*, **22**, 3574–3577.
6 Pearson,W.R. and Lipman,D.J. (1988) *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
7 Bairoch,A., Bucher,P. and Hofmann,K. (1997) *Nucleic Acids Res.*, **24**, 217–221.
8 Hirokawa,T., Boon-Chieng,S. and Mitaku,S. (1998) *Bioinformatics*, **14**, 378–379.