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

Reiko Kikuno*, Takahiro Nagase, Mina Waki and Osamu Ohara

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

Received September 17, 2001; Accepted September 18, 2001

ABSTRACT

We have been developing a HUGE database to summarize results from the sequence analysis of human novel large (>4 kb) cDNAs identified in the Kazusa cDNA sequencing project, systematically designated KIAA plus a four-digit number. HUGE currently contains nearly 2000 gene/protein characteristic tables harboring the results of the computer-assisted analysis of the cDNA and the predicted protein sequences together with those of expression profiling and chromosomal mapping. In the updated version of HUGE, we made it possible to compare each KIAA cDNA sequence with the corresponding entry in the human draft genome sequence that was published recently. Approximately 90% of KIAA cDNAs in HUGE can be localized along the human genome for at least half or more of the cDNA's length. Any nucleotide differences between the cDNA and the corresponding genomic sequences are also presented in detail. This new version of HUGE greatly helps us evaluate the completeness of cDNA clones and the accuracy of cDNA/genomic sequences. More interestingly, in some cases, the ability to compare cDNA with genomic sequences allows us to identify candidate sites of RNA editing. HUGE is available on the World Wide Web at http://www.kazusa.or.jp/huge.

INTRODUCTION

We have been conducting the Kazusa cDNA sequencing project with the aim of (i) identifying previously unidentified human transcripts, (ii) predicting the primary structure of gene products encoded by the unidentified human transcripts, and (iii) characterizing the unidentified gene products. In particular, the Kazusa cDNA project has focused on long cDNA clones which direct the synthesis of large proteins (>50 kDa) (1). Nearly 2000 human cDNA sequences (average size: 4.7 kb) have been published to date (2), and the results of the sequence analysis, expression profiling and chromosomal mapping for respective cDNAs have been summarized in the HUGE database (3).

In the post-genomic era, we need a set of cDNA clones with the capability of producing functional proteins. However, it is not an easy task to make all the cDNA clones expression-ready because cDNA is only an artificial DNA copy of poly(A)-tailed mRNA and thus the clones are not completely free from artificial errors in their sequences. In this respect, large cDNAs as compared to short cDNAs generally have a higher risk of containing artifacts, which may be caused by reverse transcriptase error and/or intronic sequence(s) accidentally retained in the template mRNA. To avoid incorrectly predicting the protein-coding sequences in KIAA cDNAs, we examined all the KIAA sequences in terms of protein-coding potentiality by GeneMark analysis, and when a warning for spurious coding interruption was triggered, we performed additional experiments using the reverse transcription-coupled polymerase chain reaction method to detect artifacts and correct the sequence data (4). Detailed results are presented in HUGE.

Evaluating the completeness of the cDNA clones in terms of their 5′- and 3′-termini is also important and can be accomplished by localizing cDNA sequences on the genome. By comparing the cloned cDNA sequences and the transcribed sequences as predicted by a gene-finding program, the integrity of 5′- and 3′-termini as well as the possible presence of retained introns in the cloned cDNA can be studied. For this purpose, we mapped each cDNA sequence on the human genome draft sequence published this year (5), then applied the GENSCAN program (6) to the genomic region where the cDNA was assigned. The results of the comparison between the cloned and predicted structures are graphically presented in HUGE. In addition, we also describe the extent of nucleotide identity between the cDNA and genomic sequences. Using these types of information as a guide, we have been updating our cDNA sequence data after experimental evaluation when necessary, to achieve a complete and accurate catalog of human large cDNAs.

ORGANIZATION OF THE HUGE DATABASE

The format of cDNA entries in the HUGE database has been designated as KIAA plus a four-digit number, e.g. KIAA0001. Each entry has its own gene/protein characteristic table showing the cDNA sequence, results from the computer analyses of the sequence at both the DNA and protein levels, chromosomal locations and expression profiles, as described previously (3). Some new features in each gene/protein characteristic table are described below.

*To whom correspondence should be addressed. Tel: +81 438 52 3932; Fax: +81 438 52 3931; Email: kikuno@kazusa.or.jp

EXAMINATION OF THE CODING POTENTIALITY BY GeneMark ANALYSIS

We examined all the sequences for protein-coding potentiality by GeneMark analysis as described previously (3). When GeneMark analysis triggered an alert for spurious coding interruptions, we performed additional experiments using the reverse transcription-coupled polymerase chain reaction method to detect artifacts and correct the sequence data. The number of such revisions carried out to date is 262, constituting 13.5% of all the cDNA sequences in HUGE. We have added new detailed information about this process to HUGE, in addition to the comparative alignment of the cloned and the revised sequences, to avoid unnecessary confusion.

PREDICTED GENOMIC STRUCTURE OF KIAA GENES

Entries of the human draft genome sequences were downloaded from the NCBI ftp site (ftp://ncbi.nlm.nih.gov/genomes/H_sapiens/). Our standard procedure was to first search the genome contigs using BLAST for a highly similar segment to the KIAA cDNA sequence under study (7). When a highly similar fragment (*E*-value = 0.0 and sequence identity 90% or more) was successfully found in a particular genome contig, we then applied the SIM4 program to assign the cDNA sequence to the genome contig (8). The allocation results of a cDNA sequence along a corresponding genome contig are summarized in a table that presents the nucleotide positions in the genome contig, the position of the protein-coding region, the identity between the cDNA and genomic sequences for each exon, and the boundary dinucleotide sequences and lengths of the predicted introns. We also applied the GENSCAN program to the genome contig to predict plausible KIAA gene structure(s). The exon probabilities calculated by GENSCAN are shown in color gradations so that it is possible to compare directly the predicted transcript structure with the cloned KIAA cDNA sequence.

HUGE contains 1934 KIAA cDNA entries at present, of which 1743 cDNAs (90%) can be aligned for over half of their cDNA lengths. While 78 KIAA cDNAs (4%) do not have any corresponding genomic entries at all, 1152 cDNAs (60%) can be perfectly assigned on the genome. This assignment rate is slightly higher than that of RefSeq (85%) of NCBI, a database of the manually curated collection of most full-length human mRNA sequences in GenBank (5,9). In this regard, HUGE is thus considered one of the most reliable collections of mRNA sequences. When multiple genomic entries are detected for single KIAA genes, HUGE provides all the possible mapping data of cDNAs on the genome.

NUCLEOTIDE DIFFERENCES BETWEEN THE cDNA AND GENOME SEQUENCES

cDNA sequences are not always perfectly identical with their corresponding genomic sequences even if the cDNA can be assigned to a specific area of the genome over the entire region. There may be several reasons for these discrepancies: (i) incorrect genome mapping due to the recent segmental duplication of the genome, (ii) polymorphisms, (iii) the presence of remaining gap(s) in the genome contigs, (iv) error in either the genome or cDNA sequencing, (v) error in reverse transcription, and (vi) RNA editing. We cannot easily determine which of these reasons applies to each case, but it is critical to be aware of the presence of nucleotide differences between the cDNA and genomic sequences upon utilization of cDNA clones. Taking this into consideration, we constructed tables of nucleotide substitutions and insertions/deletions found between each KIAA cDNA sequence and the corresponding genome sequence.

A prominent feature of the nucleotide substitutions between the cDNA and genomic sequences is the predominance of clustered A-to-G changes (from genome to cDNA), which is difficult to explain by single nucleotide polymorphism, reverse transcription error and/or sequencing error. This is a reminder that the brain, from which most of KIAA cDNAs were derived, has been reported to be rich in adenosine deaminase activity specific to double-stranded RNA (ADAR) (10). ADAR is known to convert adenosine to inosine if adenosine is present in double-stranded RNA and the resultant inosine residue is reverse-transcribed to G in cDNA (11). Thus, the action of ADAR may be a plausible explanation for the predominance of A-to-G changes in KIAA cDNAs. When we examined the observed A-to-G changes more closely, it was observed that the A-to-G changes were frequently clustered in a relatively short region of specific KIAA cDNAs. Interestingly, most of the clustered A-to-G changes are located in Alu elements. In particular, the clustered A-to-G changes in single KIAA cDNAs were frequently observed in regions including two Alu elements inversely oriented. Because the inversely oriented two Alu elements are very likely to form a double-stranded intrastructure in the mRNA, this observation is consistent with the assumption that the clustered A-to-G changes resulted from the action of ADAR. The section of the table illustrating 'Nucleotide differences between the cDNA and the genome' may thus provide interesting information on the RNA editing of KIAA genes.

When only a single part of the genome is assigned for the entire region of the cDNA and contains a single-base nucleotide difference(s) that interrupts the coding sequence, a mistake in mapping, a polymorphism or the presence of a remaining gap(s) are not likely to be reasons for the discrepancy. We found that 323 KIAA cDNAs in HUGE were assigned to genomic regions apparently possessing a non-sense mutation or a frame-shift mutation, while 292 KIAA cDNAs exhibited sequence identity >99% to the genomic sequence in their entire length. The frequency of the former type of genomic sequences [323/1152 (28%) of KIAA cDNAs entirely assigned to the genome] was too high to be explained by errors in reverse transcription, RNA editing and/or cDNA sequencing. Although the cause of the interruption in coding-sequences that are only on the genome cannot be conclusively determined at present, this observation is a caution when using the human draft genome sequence for *in silico* gene prediction.

EXTERNAL LINKS TO THE OTHER DATABASES

HUGE has mutual links to three other databases, GTOP (http://spock.genes.nig.ac.jp/~genome/gtop.html), SWISS-PROT/TrEMBL and GeneCards (12,13). Users can obtain the predicted 3D structures of the protein products through GTOP, the results of other various motif/domain search through SWISS-PROT/TrEMBL, and information on the involvement in genetic disease through GeneCards. These links provide many benefits to users of HUGE because they can retrieve a wide variety of information regarding KIAA genes cataloged in HUGE.

FUTURE DIRECTIONS

Although the human draft genome sequence has become publicly available, it is expected that the complete catalog of human proteins will take a some time to establish. Because large transcripts are difficult targets for structural analysis in general, HUGE serves as a unique informational platform for human protein analysis. As discussed previously, we will continuously update HUGE and, if necessary, revise our cDNA sequences as well as integrate new types of information on KIAA cDNAs. As one of the fundamental bases of understanding the human proteome, HUGE will become a more integrative database which will enable an overview not only of sequence information but also of functional information obtained by various experimental approaches, such as DNA microarray and protein–protein interaction assay.

ACKNOWLEDGEMENTS

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) Construction and characterization of human brain cDNA libraries suitable for analysis of cDNA clones encoding relatively large proteins. *DNA Res.*, **4**, 53–59.
2. Nagase,T., Nakayama,M., Nakajima,D., Kikuno,R. and Ohara,O. (2001) Prediction of the coding sequences of unidentified human genes. XX. The complete sequences of 100 new cDNA clones from brain which code for large proteins *in vitro*. *DNA Res.*, **8**, 85–95.
3. Kikuno,R., Nagase,T., Suyama,M., Waki,M., Hirosawa,M. and Ohara,O. (2000) HUGE: a database for human large proteins identified in the Kazusa cDNA sequencing project. *Nucleic Acids Res.*, **28**, 331–332.
4. Hirosawa,M., Isono,K., Hayes,W. and Borodovsky,M. (1997) Gene identification and classification in the *Synechocystis* genomic sequence by recursive gene mark analysis. *DNA Seq.*, **8**, 17–29
5. International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
6. Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.
7. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
8. Florea,L., Hartzell,G., Zhang,Z., Rubin,G.M. and Miller,W. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.*, **8**, 967–974.
9. Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137–140
10. Paul,M.S. and Bass,B.L. (1998) Inosine exists in mRNA at tissue-specific levels and is most abundant in brain mRNA. *EMBO J.*, **17**, 1120–1127
11. Bass,B.L. (1997) RNA editing and hypermutation by adenosine deamination. *Trends Biochem. Sci.*, **22**, 157–162.
12. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
13. Rebhan,M., Chalifa-Caspi,V., Prilusky,J. and Lancet,D. (1998) GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics*, **14**, 656–664.