Genomics & Informatics Vol. 3(3) 86-93, September 2005

KBUD: The Korea Brain UniGene Database

Yeo-Jin Jeon[1], Jung-Hwa Oh[1], Jin-Ok Yang[2] and Nam-Soon Kim[1],*

[1]Laboratory of Human Genomics, Genome Research Center, [2]National Genome Information Center (NGIC), Korea Research Institute of Bioscience and Biotechnology (KRIBB), Daejeon 305-333, Korea

Abstract

Human brain EST data provide important clues for our understanding of the molecular biology associated with the function of the normal brain and the molecular pathophysiology with brain disorders. To systematically and efficiently study the function and disorders of the human brain, 45,773 human brain ESTs were collected from 27 human brain cDNA libraries, which were constructed from normal brains and brain disorders such as brain tumors, Parkinson's disease (PD) and epilepsy. An analysis of 45,773 human brain ESTs using our EST analysis pipeline resulted in 38,396 high-quality ESTs and 35,906 ESTs, which were coalesced into 8,246 unique gene clusters, showing a significant similarity to known genes in the human RefSeq, human mRNAs and UniGene database. In addition, among 8,246 gene clusters, 4,287 genes (52%) were found to contain full-length cDNA clones. To facilitate the extraction of useful information in collected these human brain ESTs, we developed a user-friendly interface system, the Korea Brain Unigene Database (KBUD). The KBUD web interface allows access to our human brain data through three major search modes, the BioCarta pathway, keywords and BLAST searches. Each result when viewed in KBUD offers comprehensive information concerning the analyzed human brain ESTs provided by our data as well as data linked to various other public databases. The user-friendly developed KBUD, the first world-wide web interface for human brain EST data with ESTs of human brain disorders as well as normal brains, will be a helpful system for developing a better understanding of the underlying mechanisms of the normal brain well as brain disorders. The KBUD system is freely accessible at http://kugi.kribb.re.kr/KU/cgi-bin/brain.pl.

*Corresponding author: E-mail nskim37@kribb.re.kr, Tel +82-42-879-8112, Fax +82-42-879-8119
Accepted 31 August 2005

Keywords: KBUD, human brain ESTs, full-length cDNA, brain database

Introduction

The large-scale EST collections of mammalian brains including the human brain were performed to rapidly identify the expressed genes and understand the gene regulation in the brain. These brain ESTs have been produced as the result of worldwide efforts, with major contributions being made through the Brain Molecular Anatomy Project (BMAP, http://trans.nih.gov/bmap/), the NIH-Mammalian Gene Collection (MGC, http://mgc.nci.nih.gov/), and the Cancer Genome Anatomy Project (CGAP, http://cgap.nci.nih.gov). These efforts permit the discovery of transcripts that are expressed in the brain and attempt to localize the site of expression of all transcripts in the brain by large-scale parallel analyses of gene expression. In addition, the collected brain ESTs can serve as key resources for brain functional studies based on genomics, including exon detection, alternative splicing analysis, single-nucleotide polymorphisms (SNPs) and gene expression studies using a microarray as well as gene discovery and gene mapping (Adams *et al.*, 1992; Boguski *et al.*, 1995; Ermolaeva *et al.*, 1998; Mathe *et al.*, 2002; Picoult-Newberg *et al.*, 1999). In addition, an EST set obtained from specific regions of the brain in the case of certain brain diseases provides molecular insights into biological phenomena and cellular physiology under specific brain disease conditions.

Several databases for genes that are expressed in the brain have been reported to date. The NCBI's dbEST database (http://www.ncbi.nlm.nih.gov/dbEST/) supplies BMAP mouse brain EST sequences. The CGAP website provides an index of the genes that are generated in various cancer tissues including brain cancer in humans and the mouse. In addition, the Electronic Atlas of the Developing Human Brain (EADHB) database (http://www.ncl.ac.uk/ihg/EADHB/database/EADHB_database.html) and the EMAGE Mouse Gene Expression Database provide gene expression patterns during the early development of the human brain, and information related to spatial 2D and 3D visualizations for them. The human brain EST database, which provides information concerning

genes expressed in normal human brain as well as human brain disorders such as brain tumor and PD, has not been reported, although the BMAP database supplies mouse brain ESTs and the CGAP database supplies for genes expressed in various cancer tissues including human brain cancer.

To systematically and efficiently study the function and disorders of the human brain, a collection of an entire set of genes expressed in the normal brain and in brain disorders such as brain tumor, PD and epilepsy was performed as a part of the MOST 21C Frontier R & D program in neuroscience, started in Korea. For the efficient collection of full-length cDNAs and rarely expressed genes, we used oligo-capping methods to construct full-length cDNA libraries (Oh *et al.*, 2004) and Bento Soares's method to construct a normalized cDNA library (Soares *et al.*, 1994). By large scale sequencing of the clones from these brain libraries, a large number of human brain ESTs could be generated. To extract various type of information from these brain ESTs and facilitate their use, we developed a user-friendly web interface, the Korea Brain Unigene Database (KBUD) system. In this paper, we describe the human brain EST analysis pipeline, the construction of the KBUD system and its use. The KBUD system provides the various information of our human brain ESTs and allows the investigator to undertake a sophisticated search using relevant linked websites.

Methods

Human brain EST sequence data

The human brain EST sequences used in this study were obtained by the 5'-end sequencing of the clones from 27 high-quality brain cDNA libraries containing a full-length cDNA library, normalized cDNA libraries and subtracted cDNA libraries, constructed using various methods, as described previously (Kim *et al.*, 2004; Oh *et al.*, 2004; Soares *et al.*, 1994). The brain tissue samples used were as follows: One normal brain tissue, ten different brain tumor tissues, and one temporal lobe and three different hippocampuses from epilepsy patients as well as one normal substantia nigra from a normal subject and one substantia nigra from a PD patient.

Human brain EST analysis

Pre-processing: Base-calling and quality assessment were performed with the phred program (Ewing *et al.* 1998). Vector and linker sequences were located using the FASTA (http://www.ebi.ac.uk/fasta/) program and were subsequently removed. Low-quality bases defined by the phred program were trimmed from both ends of the EST sequences. Human repetitive elements and low complexity regions were masked using the RepeatMasker (http://repeatmasker.genome.washington.edu) program. EST sequences having an identity of at least 90% with human mitochondrial DNAs or ribosomal RNAs over at least 90bp were also excluded. ESTs of at least 100 bp after both vector and low-quality trimming were regarded as "high-quality" ESTs.

Annotation: The annotation of the ESTs was carried out by stand-alone BLAST programs and locally installed databases (Altschul *et al.*, 1990). The individual high-quality ESTs were searched against the human RefSeq (Pruitt *et al.*, 2005). The remaining ESTs were searched against the human mRNA (ftp://ftp.ncbi.nih.gov/genbank/) subset extracted from the GenBank database and subsequently against the UniGene database (Hs.seq.all, build #184) for similarity comparisons using BLASTN. We carried out a BLAST search with a cut-off identity of 97% and an E value of $1e^{-10}$, and over at least 100 bp of ESTs compared with the RefSeq, GenBank mRNA and UniGene data. After the ESTs were clustered into gene indices by the BLAST results and a cDNA clone containing the best hit score in each cluster was invested with ID (ex.BKU000001), selected as a representative unique gene. The CAP3 program was used to assemble the ESTs in each cluster into contigs (Huang *et al.*, 1999).

Determination of full-length cDNA: The fullness of our human brain ESTs was determined by comparing EST sequences against the coding sequences of Human RefSeq entries and known human mRNAs. We categorized them into four groups, as follows: 1) full-length cDNA (Full); at least 97% identity over the first 100 bases of the CDS (coding sequence), including the ATG initiation codon, 2) candidate full-length cDNA (Can-Full); matched at least 100 bp against the 5' UTR of the upstream CDS region, 3) partial cDNA (Partial); not containing an ATG initiation codon and matched at least 100 bp against CDS region, 4) 3' UTR; not covering any CDS region and matched to the 3' UTR of the downstream CDS region, 5) Unknown; matched only against ESTs or a gene having an unknown CDS region which can not be defined.

Construction of the KBUD system

KBUD was written using python and perl script for the EST annotation work and was implemented in a relational database structure using MySQL DBMS. The web-front end was handwritten in HTML. Web-based database searching was implemented with CGI-Perl scripts. To allow a BLAST search of the user's query data, the BLAST

suite of programs were downloaded from the NCBI BLAST ftp site (ftp://ftp.ncbi.nih.gov/blast/).

Results and Discussion

Generation of human brain ESTs by large-scale sequencing

For the collection of ESTs expressed in the human brain, we constructed 27 brain cDNA libraries from various human brain tissues, as shown in Table 1. Brain tumor tissues such as, glioblastoma, oligodendroglioma, neuroblastoma and meningioma, were used in order to study the pathogenesis of brain tumors. In addition, brain tissues obtained from patients with PD and epilepsy were used in order to study other brain disorders. 27 cDNA libraries consisting of 19 full-length enriched cDNA libraries, and 7 normalized cDNA libraries and 1 subtracted library were constructed from the start full-length cDNA libraries in order to discover genes that are rarely expressed in the brain. In total, 45,773 clones were randomly selected from these 27 libraries and were used for 5'-end single-pass sequencing. These brain ESTs served as the key data set of KBUD. In addition, human brain EST will continue to be generated from various cDNA libraries and will be updated in the KBUD accordingly.

Table 1. Summary of constructed human brain cDNA libraries

Tissue	No. of Library	Library Type[a]		
		Full	Nor.	Sub.
Normal	8	5	3	
Disease				
Parkinson's Disease	2	1	1	
Epilepsy	3	3		
Tumor				
Gliblastoma	8	4	3	1
Oligodendroglioma	4	4		
Meningioma	1	1		
Neuroblastoma	1	1		
Total	27	19	7	1

[a] Library type: Full, full-length enriched; Sub, subtracted; Nor, normalized.

Human brain EST analysis pipeline

The obtained 45,773 EST sequences were subjected to quality control procedures, as shown in pre-processing diagram in Fig. 1, namely, trimming of the vector region and the removal of low-quality or short (less than 100 bp) sequences. After screening out 1,616 ESTs derived from

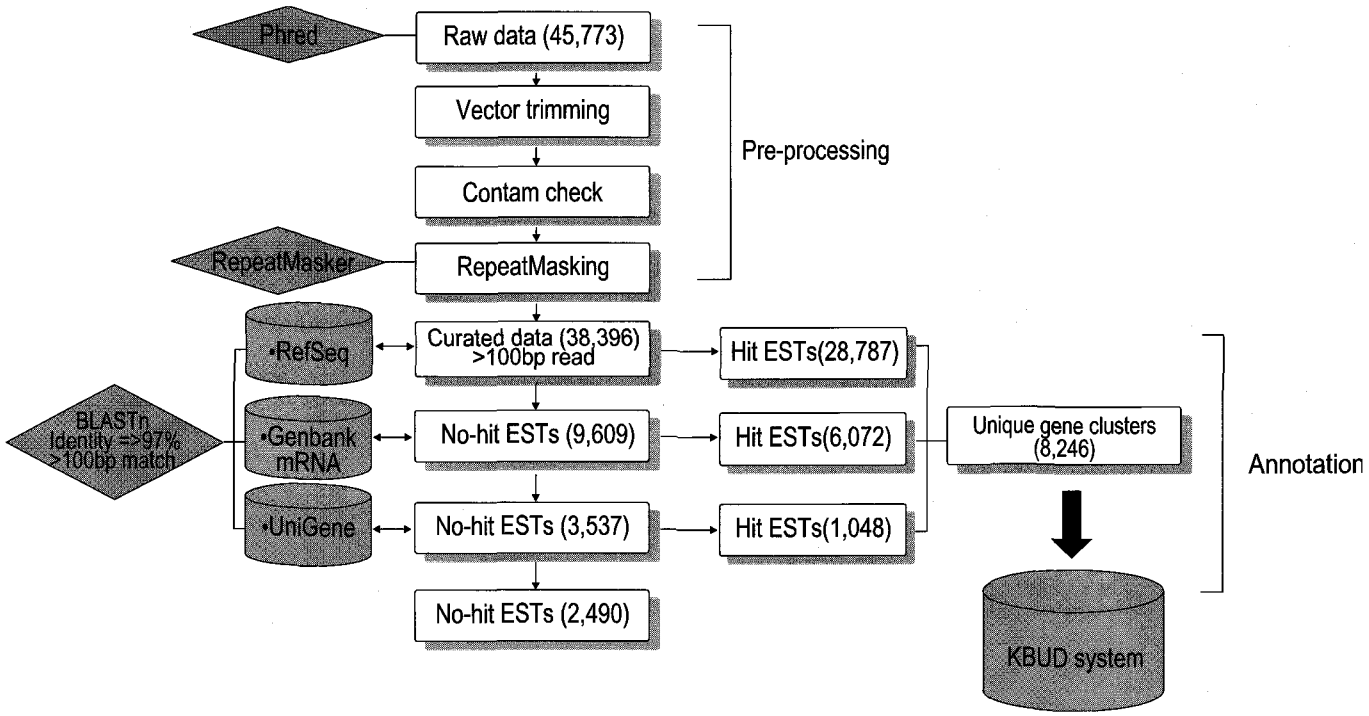

Fig. 1. Flow chart for analyzing human brain ESTs. Human brain ESTs were pre-processed and annotated using this procedure. Detailed cut-off values are described in the method section. The number of ESTs and unique gene clusters in each category are represented in parenthesis.

mitochondrial DNAs, ribosomal DNAs, and human repetitive sequences, 38,396 high-quality ESTs with an average length of 589 bp were generated.

To annotate high-quality ESTs, 38,396 ESTs were sequentially analyzed, as shown in the annotation process of Fig. 1. When we compared these ESTs with human RefSeq using the BLAST program, 28,787 ESTs (75%) showed a significant similarity to known genes in the human RefSeq and the remaining 6,072 ESTs (15%) were matched to human mRNAs. In addition, the remaining 1,048 EST (3%) were matched to UniGene ESTs and 2,490 (7%) showed no match or a match lower than 97% nucleotide identity to previously reported known genes in the public database. A total of 35,906 known ESTs were clustered into 8,246 unique known genes based on BLAST similarity scores. The cluster number per total known ESTs was estimated to be about 23% (8,246 /35,906) from a transcript analysis using the CAP3 program. The complexity of the collected genes, on average, was slightly higher than that of the Cap-trapper cDNA libraries collated by normalization and subtraction (Carninci *et al.*, 2000). Table 2 shows that 8,246 unique gene clusters consisting of 4,777 of contigs and 6,597 of singletons were obtained using the CAP3 program. The 4,777 contigs were assembled from 29,309 EST sequences.

To determine the fullness of our human brain ESTs, 8,246 gene clusters for 38,396 high quality ESTs were compared to human RefSeq entires and known mRNA. As shown in Fig. 2, 4,287 genes (52%) were shown to contain full-length cDNA clones (21,434), whereas 291 (4%) contained candidate full-length cDNA clones (1,198). In addition, 1,908 genes (23%) contained only partial cDNA clones (7,062) and 1,760 (15%) matched only the 3' UTR of genes from cDNA clones (5,164). The frequency of full-length cDNA containing full-length

Table 2. Human brain EST clustering results

	No
Total sequencing data	45,773
Pre-processed ESTs	40,012
Mitochondrial DNA and others	1,616
High-quality ESTs	38,396
Unique known genes	8,246
Contigs	4,777
Singletons	6,597
ESTs in contigs	29,309
No-hit ESTs	2,409

For unique known genes, contigs and singlets were counted as clusters using CAP3 program.
The ESTs clones were clustered by BLASTN search results against RefSeq, human mRNA and UniGene databases.
Each definition was described in detail in the methods section.

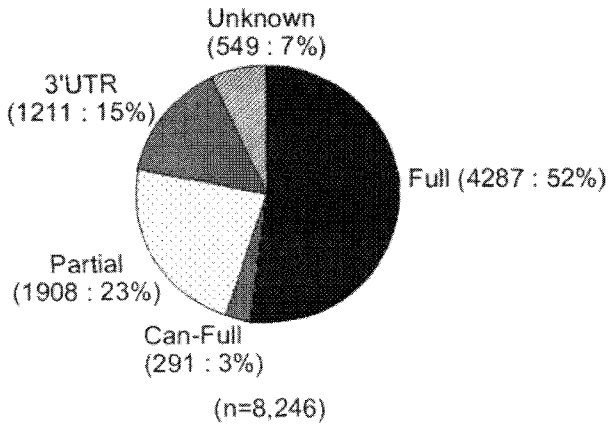

Fig. 2. Analysis of full-length cDNAs from human brain ESTs. The contents of the full-length cDNA in the unique gene clusters were represented as a pie-shaped graph. Full, full-length cDNA; Can-Full, candidate full-length cDNA; Partial, partial cDNA; 3' UTR, cDNA containing only 3' UTR; Unknown, matched only against ESTs or gene having an unknown CDS region. All of the definition was described detailly in the methods section.

cDNA and candidate full-length cDNA is 56%, slightly lower than that reported by other researchers using full-length cDNA libraries. These data concerning full-length cDNAs generated from various brain sources help in understanding the regulation of the transcription of genes expressed in the human brain and could be used in a functional analysis of the human brain. All of these data are stored in the KBUD system.

KBUD system

To facilitate the use of the analyzed human brain ESTs and to provide useful information concerning these ESTs for other investigators, we constructed a user-friendly web interface, the KBUD system using our human brain EST data set. As shown in Fig. 3, the KBUD web interface allows easy access to the data through three major search modes, the BioCarta pathway, keyword and BLAST searches. Each result provides concise information concerning an interesting gene in the human brain ESTs and more detailed information can be obtained using data linked to diverse public databases such as Unigene, OMIM and Ensemble. The public database links and sequence analysis links provide a variety of resources related to Korea Brain UniGene (KBU) on the internet. The wide variety of information for KBU using useful links permit the user to more easily and quickly analyze the gene for further study.

Usage of KBUD system

From our KBUD web page at http://kugi.kribb.re.kr/KU/

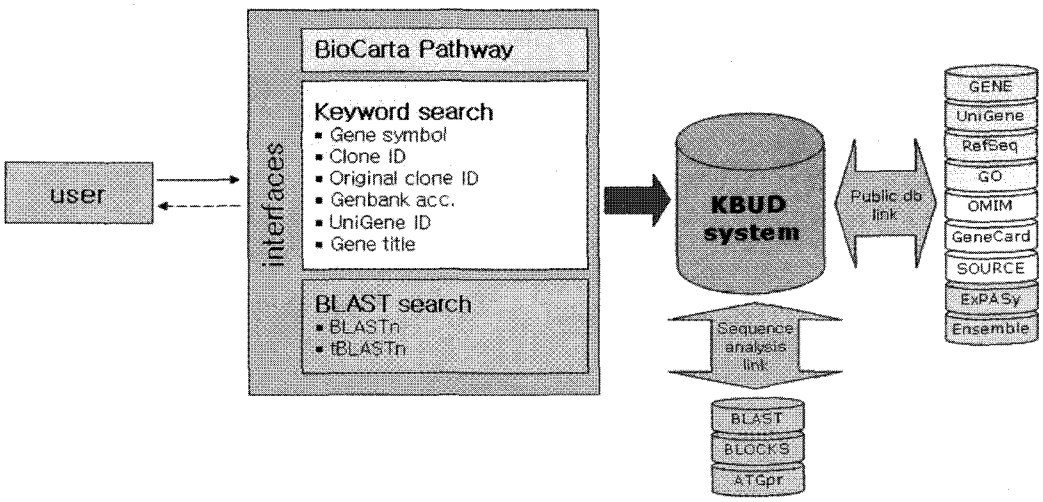

Fig. 3. System architecture of the KBUD system. User can access the KBUD web interface through three major search modes such as BioCarta pathway, keyword and BLAST search and then obtained concise information about the interesting Korea Brain UniGene (KBU) which are provided by our data as well as diverse public database.

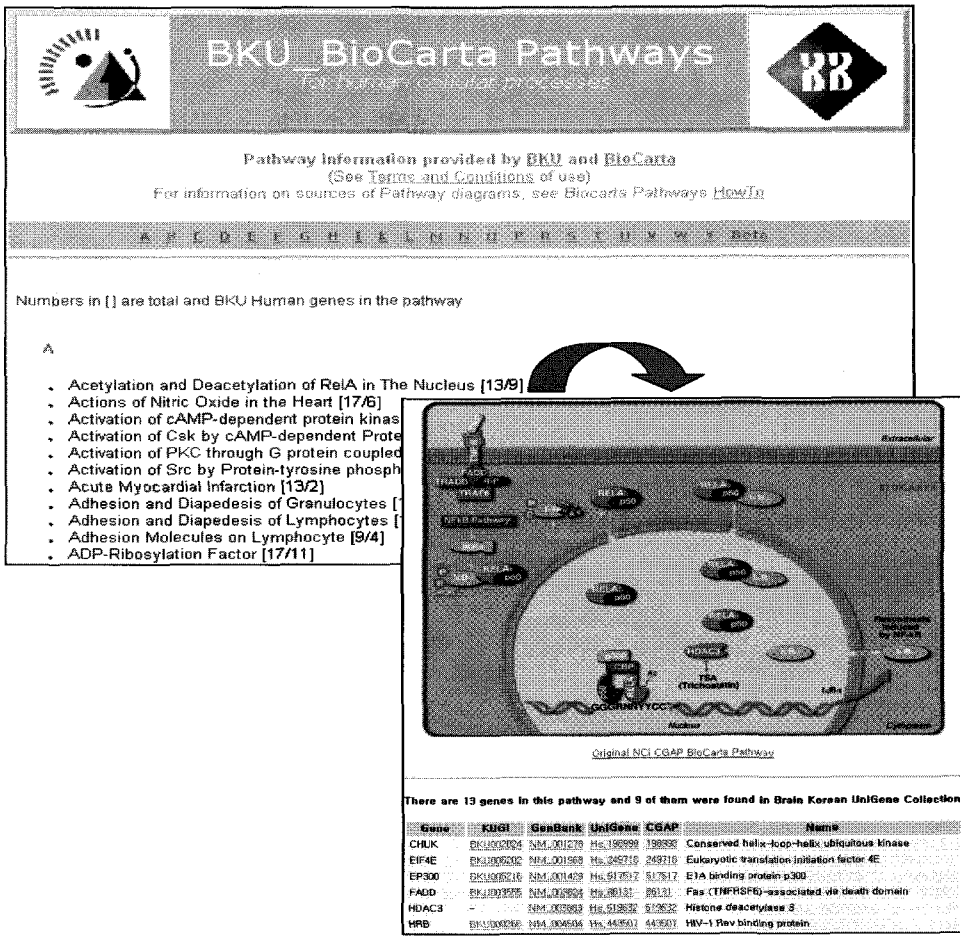

Fig. 4. KBUD information obtained through the BioCarta pathway viewer. This view offers a dynamic graphical model for 311 pathways and various information on KBU involved in these pathway.

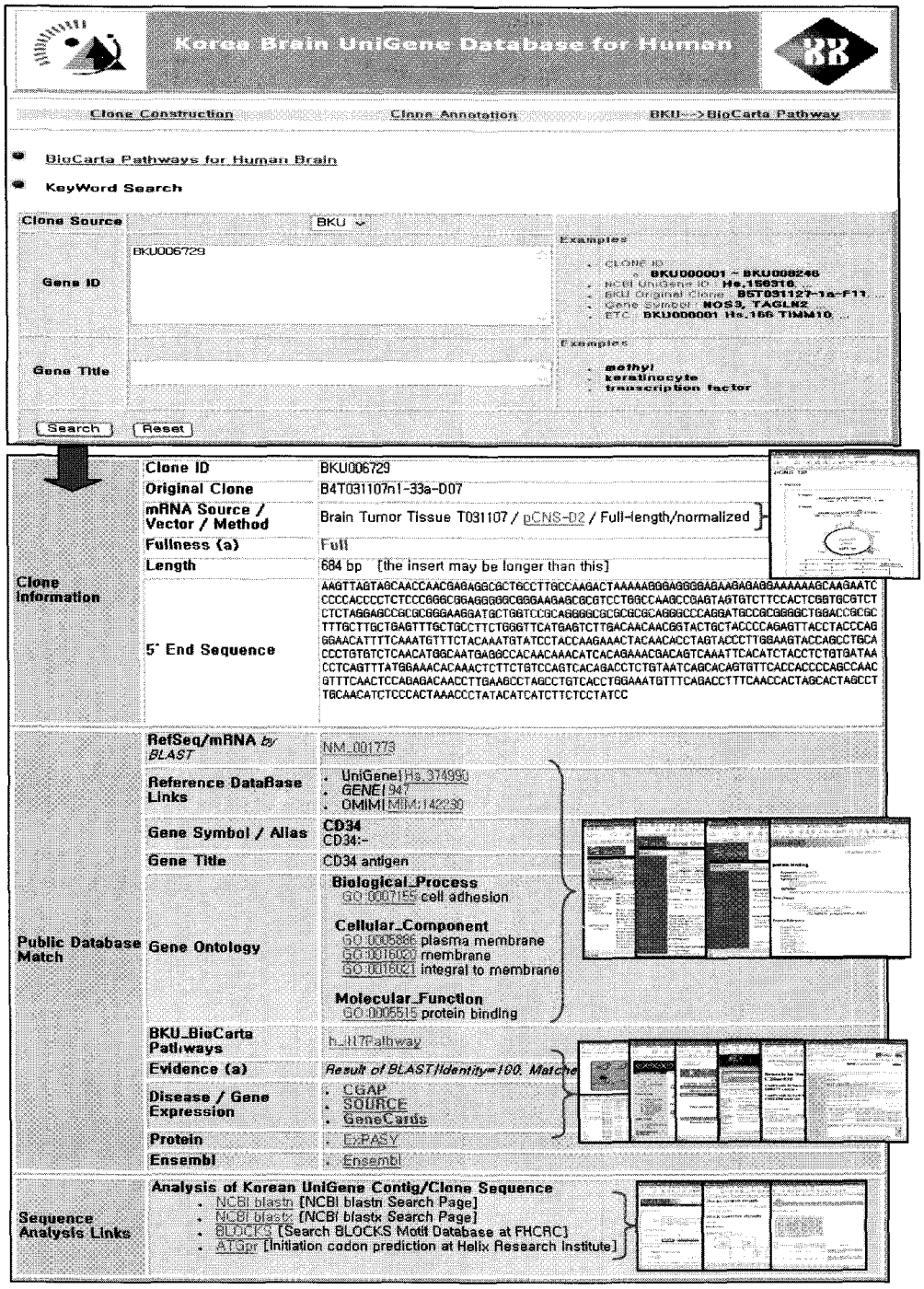

Fig. 5. KBUD information through a keyword search viewer. This view offers various KBU clone information which consists of clone information, a public database match and sequence analysis links.

cgi-bin/brain.pl, users can retrieve various information on human brain ESTs in several ways as described below.

In the pathway search viewer, displays of protein-protein interactions within pathways for human cellular processes are provided, including diagrams of molecular assemblies. Users can select a pathway of interest from a list of 311 pathways which are arranged in alphabetical order, and the result view offers a dynamic graphical model and the summarized table of genes involved in the pathway, as shown in Fig. 4. The graphical model shows how genes

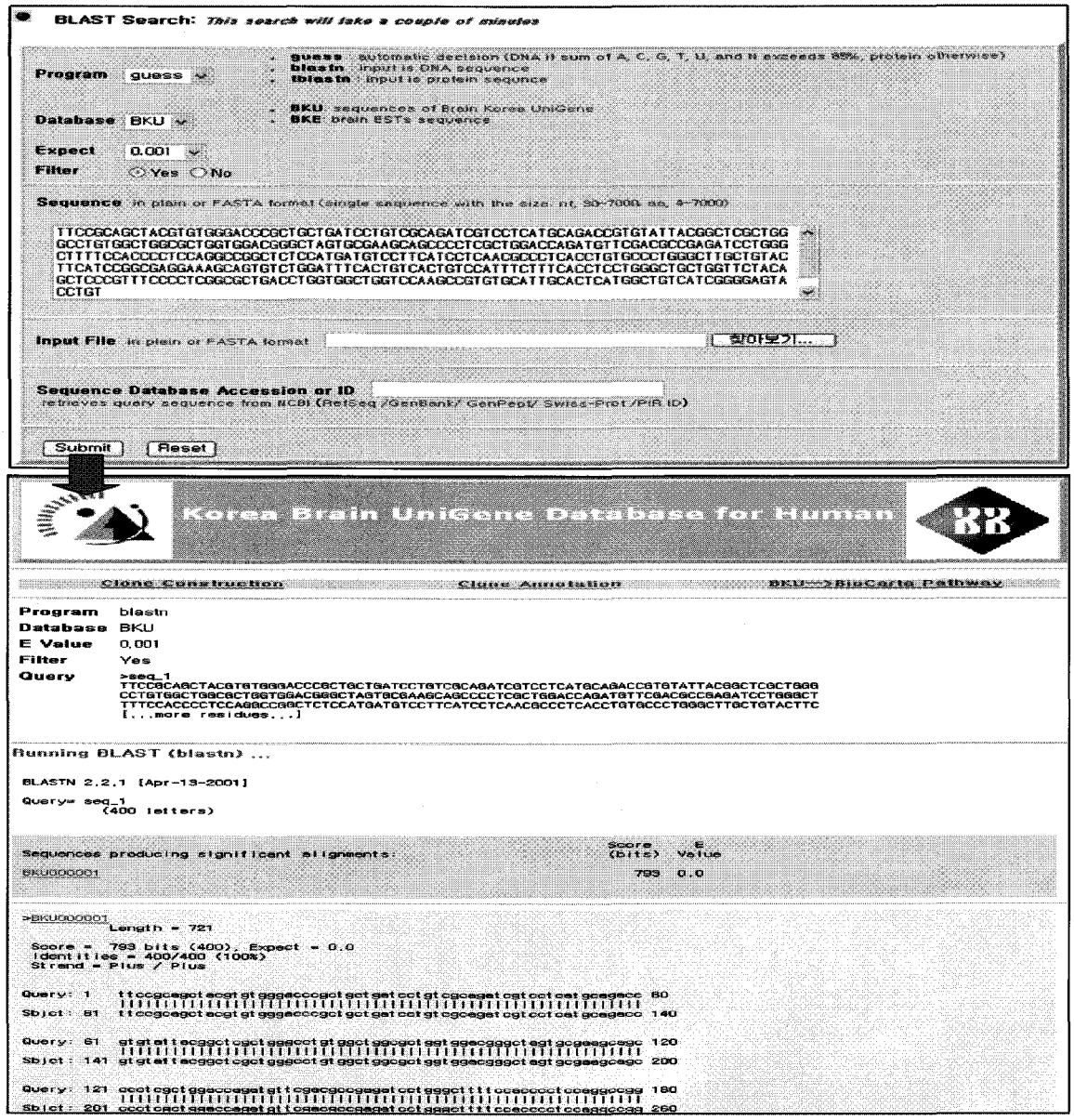

Fig. 6. KBUD information through a BLAST search. This viewer allows the BLAST searching of the user's query against our KBU data set.

interact in the pathway and the summarized table provides the specific gene ID, BKU ID that represents our brain cDNA. Each specific gene ID is directly linked to its gene information page generated from several public databases such as GenBank, UniGene and CGAP. The pathway search viewer allows the user to easily access entire sets of interacting genes in a specific pathway. The information can be used to facilitate an understanding of the molecular pathway and to help design more focused experiments for validating the biological process in human brain disease and in therapeutics.

Through the keyword search viewer, users can also search for interesting brain cDNAs using the gene symbol, BKU ID, original clone ID, Genbank accession number, UniGene ID and gene title. Fig. 5 shows the output image screen of keyword search viewer. The result view shows the summarized gene information of the brain cDNA of interest, KBU, and provides the sophisticated information derived from other linked useful websites. This page consists of a three-party summarized table such as clone information, public database match and sequence analysis links. The clone information part provides basic brain

cDNA data such as the mRNA source, vector, fullness, 5' end sequence and its length read, supplied by us. The public database match shows various type of information on annotated gene generated from matched RefSeq or mRNA, OMIM, Gene Ontology (GO), BioCarta pathway database and BLAST result. Moreover, Links to other well-known databases provide information concerning related diseases or gene expression and protein structure or genome information. The sequence analysis part also offers a useful website such as BLOCKS and ATGpr to search the motif or initiation site of the sequence.

In addition, the BLAST search viewer allows BLAST searching of the user's query against our brain data set through a user-friendly implemented interface with the BLAST programs such as, BLASTN, TBLASTN, as shown in Fig. 6.

Information on brain ESTs as the result of large scale analysis is known to provide important clues for understanding the functions of the nervous system and the molecular pathophysiology of brain disorders (Boguski and Jones, 2004; Gong *et al.*, 2003). KBUD is the first world-wide web interface for human brain EST data generated mainly from Korean patients with brain disorders such as brain tumors, PD and epilepsy. The KBUD system is quickly and easily accessible to any investigator for searching the interesting brain cDNAs and also permits more detailed analyses using useful linked websites in our system. The user-friendly developed KBUD system will be a helpful system in understanding brain tumor biology or neurodegenerative pathology.

Acknowledgements

This work was supported by a grant M103KV01000303 K2201 of the MOST 21C Frontier R & D program in neuroscience from the Ministry of Science & Technology of Korea. We thank Dr. Sangsoo Kim in KRIBB for technical support and advices in database construction. We thank Ju-Yeon Lee, Sang-Soon Byun, Tae-Hoon Song and Sin-Ae Lee in KRIBB for their excellent technical support in the EST sequencing.

References

Adams, M.D., Dubnick, M., Kerlavage, A.R., Moreno, R., Kelley, J.M., Utterback, T.R., Nagle, J.W., Fields, C., and Venter, J.C. (1992). Sequence identification of 2,375 human brain genes. *Nature* 355, 632-634.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403-410.

Boguski, M.S. and Schuler, G.D. (1995). ESTablishing a human transcript map. *Nat. Genet.* 10, 369-371.

Boguski, M.S. and Jones, A.R. (2004). Neurogenomics: at the intersection of neurobiology and genome sciences. *Nat. Neurosci.* 7, 429-433.

Carninci, P., Shibata, Y., Hayatsu, N., Sugahara, Y., Shibata, K. Itoh, M., Konno, H., Okazaki, Y., Muramatsu,M., and Hayashizaki, Y., (2000). Normalization and subtraction of Cap-trapper-selected cDNAs to prepare full-length cDNA libraries for rapid discovery of new genes. *Genome Res.* 10, 1617-1630.

Ermolaeva, O., Rastogi, M., Pruitt, K.D., Schuler, G.D., Bittner, M.L., Chen, Y., Simon, R., Meltzer, P., Trent, J.M., and Boguski, M.S. (1998). Data management and analysis for gene expression arrays. *Nat. Genet.* 20, 19-23.

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* 8, 175-185.

Gong, S., Zheng, C., Doughty, M.L., Losos, K., Didkovsky, N., Schambra, U.B., Nowak, N.J., Joyner, A., Leblanc, G., Hatten, M.E., and Heintz, N. (2003). A gene expression atlas of the central nervous system based on bacterial artificial chromosomes. *Nature* 425, 917-925.

Kim, N-S., Hahn, Y., Oh, J.H., Lee, J.Y., Oh, K.J., Kim, J.M., Park, H.S., Kim, S., Song, K.S., Rho, S.M., Yoo, H.S., and Kim, Y.S. (2004). Gene cataloging and expression profiling in human gastric cancer cells by expressed sequence tags. *Genomics* 83, 1024-1045.

Mathe, C, Sagot M.F., Schiex, T., and Rouze, P. (2002). Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res.* 30, 4103-4117.

Oh, J.H., Sohn, H.Y., Kim, J.M., Kim, Y.S., and Kim, N-S. (2004). Construction of multi-purpose vectors pCNS and pCNS-D2 are suitable for collection and functional study of large-scale cDNAs. *Plasmid* 51, 217-226.

Picoult-Newberg, L., Ideker, T.E., Pohl, M.G., Taylor, S.L., Donaldson, M.A., Nickerson, D.A., and Boyce-Jacino, M. (1999). Mining SNPs from EST databases. *Genome Res.* 9, 167-174.

Pruitt, K., Tatusov, T., and Maglott, D. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 33, D501-D504.

Soares, M.B., Bonaldo, M.F., Jelene, P., Su, L., Lawton, L., and Efstratiadis A.. (1994). Construction and characterization of a normalized cDNA library. *Proc. Nati. Acad. Sci. USA* 91, 9228-9232.

Huang, X. and Madan, A. (1999). CAP3: a DNA sequence assembly program. *Genome Res.* 9, 868-877.