KUGI: A Database and Search System for Korean Unigene and Pathway Information

Jin Ok Yang[1], Yoonsoo Hahn[4], Nam-Soon Kim[2], Ungsik Yu[1], Hyun Goo Woo[1], In-Sun Chu[1], Yong-Sung Kim[2], Hyang-Sook Yoo[2], and Sangsoo Kim[3]

[1]National Genome Information Center (NGIC),
[2] The Center for Functional Analysis of Human Genome,
Korea Research Institute of Bioscience and Biotechnology (KRIBB), Daejon (305-335), Korea
[3] Department of Bioinformatics and Life Science, Soongsil University, Seoul 156-743, Korea
[4] Laboratory of Molecular Biology, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892-4264, USA
E-mail: joy@.kribb.re.kr, sskimb@ssu.ac.kr

ABSTRACT: KUGI (Korean UniGene Information) database contains the annotation information of the cDNA sequences obtained from the disease samples prevalent in Korean. A total of about 157,000 5'-EST high throughput sequences collected from cDNA libraries of stomach, liver, and some cancer tissues or established cell lines from Korean patients were clustered to about 35,000 contigs. From each cluster a representative clone having the longest high quality sequence or the start codon was selected. We stored the sequences of the representative clones and the clustered contigs in the KUGI database together with their information analyzed by running Blast against RefSeq, human mRNA, and UniGene databases from NCBI. We provide a web-based search engine for the KUGI database using two types of user interfaces: attribute-based search and similarity search of the sequences. For attribute-based search, we use DBMS technology while we use BLAST that supports various similarity search options. The search system allows not only multiple queries, but also various query types. The results are as follows: 1) information of clones and libraries, 2) accession keys, location on genome, gene ontology, and pathways to public databases, 3) links to external programs, and 4) sequence information of contig and 5'-end of clones. We believe that the KUGI database and search system may provide very useful information that can be used in the study for elucidating the causes of the disease that are prevalent in Korean.

1 INTRODUCION

Some specific cancers, such as stomach cancer and liver cancer, occur often in Korean population. In order to discover genes associated to these cancers and to develop diagnostics and therapeutics by understanding their functions, the 21st Centry Frontier Program of Human Genome Functional Analysis was launched and hosted at Korea Research Institute of Bioscience and Biotechnology (KRIBB) in 2000. The first goal of the project was to establish clone collections isolated from Korean patient samples or established cell lines. This clone set would provide valuable resources for the functional studies [1, 2]. For example, the unique non-redundant clone set would be utilized in spotting microarray chips that could provide important gene expression profiles of the cancers mentioned above. As the features spotted in the microarrays have been originated from the kind of tissues that the microarrays were going to used against, this strategy of tissue-specific microarrays would provide more informative signals than the conventional universal chips [3]. Once genes showing dysregulated expression pattern have been identified from microarray approach, the next step would be functional validation that requires expression-ready clones [4]. In this respect, cDNA microarray approach has the advantage over those based on oligo chips. For the former, the clones used in the microarray have been stocked, while the availability of these clones is not always guarantied for the latter. The analysis of gene expression profiles and the subsequent functional study or antibody development require gene annotation information of the clones. It is the purpose of the Korean unigene information (called KUGI) to provide this comprehensive information in a user-friendly manner using web technology. Here we will describe the KUGI system and the associated query methods.

2 METHODS AND DATABASE

2.1 Collection of Korean UniGene (KU) Clones

We extracted mRNAs from stomach cell lines, liver and normal stomach tissues of Korean patients. By using the oligo-capping method [5], we made more than 81 types of cDNA libraries including full-length enriched cDNA libraries, universal cDNA libraries, subtracted full-length enriched cDNA libraries, normalized universal cDNA libraries, and subtracted universal cDNA libraries [6, 7, 8].

After 5'-end sequencing of the clones isolated from these libraries, the ESTs were quality-controlled and assembled into contigs. Out of about 157,000 ESTs about 35,000 clusters were constructed and we selected a KUG clone from each contig. The KUG clone set includes some full-length clones (22895, 39.3%) and some novel genes (7179, 12.3%) that have not been reported so far. We labelled the KUG clones of the format KUNNNNNN (2 characters KU and 6 digits).

The KUGI database includes KUG clone sequences and their functional annotation information: Information

regarding the mRNA sources from tissues and cell lines, library types, vectors, and 5'-end and contig sequences. The sequences were analyzed by running BLAST against RefSeq, human mRNA, and UniGene databases from NCBI.

2.2 Clone Annotation

After both vector and low-quality trimming according to Kim et al. (2004) [9], ESTs with at least 100 bp were classified as "high-quality" ESTs. The individual ESTs were searched against the UniGene database (build 176) for similarity comparison using BLAST. The remaining EST sequences were collapsed into clusters using BLAST, and were searched against human EST sequences (est_human as of Feb 15, 2005; our own EST sequences were excluded from the database before the analysis), a non-redundant protein database, and then the human genome sequences (UCSC hg17) [10, 11, 12]. ESTs were considered as a "hit" if they shared at least a 95% identity over at least 90 bp of the DNA sequences, or matched to a protein with E value $\leq 10^{-10}$. The contig assembly was performed using the CAP3 program with the EST sequences and quality values as input data.

Full-length cDNA clones were selected by comparing our EST sequences against the coding sequences of human known mRNAs. Human RefSeq entries and known human mRNAs containing complete coding sequences were retrieved from the UniGene database (build 176). ESTs were classified as "full-length" clones if they showed at least 95% identity over the first 90 bases of the coding sequence, including ATG initiation codon. The full-length cDNA containing the longest 5'end in each cluster was selected as representative full-length cDNA. To measure the 5' UTR length distribution of our human gastric full-length cDNAs, the relative 5' UTR lengths between our full-length cDNAs and human RefSeq mRNA were analyzed using the FASTA program and were then calculated [13].

These clones were distributed by tissues as shown in below Table 1. We have gathered information of protein, diseases, gene ontology (GO), and expression from several external tools and databases [14]. Especially these each cDNA clones were compared to human genome sequences (UCSC hg17) to indicate the transcription start site and intronic positions, alternative splicing variants and characteristics of neighbor genes.

Tissues	ESTs	Genes	Fullness (%)
Stomach	22731	9042	8181(36)
Liver	10707	5547	3042(28.4)
Brain	1504	1045	481(32)
Cervix	633	511	256(40.4)
Thymus	32	25	7(21.8)

Table 1. Distribution of annotated clones

2.3 KU_BioCarta Pathways Mapping

Biological pathway mapping is an important problem in the post-genomic era [15]. Equally important and challenging as EST and genome annotation, is the subsequent classification of known or predicted genes into their respective pathways. We represent a database consisting of known genes and displays of gene interactions within pathways for human cellular processes, such as apoptosis and signal transduction [16]. To construct the KU_BioCarta Pathways, we downloaded images and gene information of pathways present in the CGAP (Cancer Genome Anotomy Project), and then mapped clones with gene name or unigene cluster on BioCarta Pathways. We can be found these clones mapped on 313 pathways and displayed on KU_BioCarta Pathways and table results.

2.4 Web-based Search Engine

The KUGI search system was implemented on the web to provide search capability of gene information as well as links to several external databases together.

The main web page of the KUGI search system is shown in the Fig.1. It contains pathway, annotation, and distribution for both Korean clones as well as mouse clones acquired from BMAP (Brain Molecule Anatomy Project) and NIA (National Institute of Aging at NIH). The page also provides a link to the clone request web page in 21C Frontier Human Gene Bank shown as Fig 10.

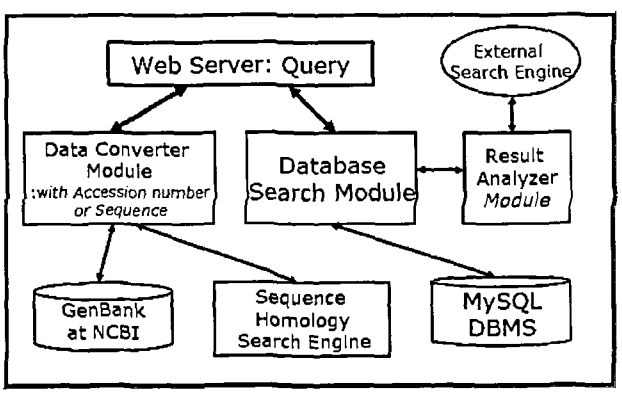

Fig. 3 Architecture of KUGI Search System

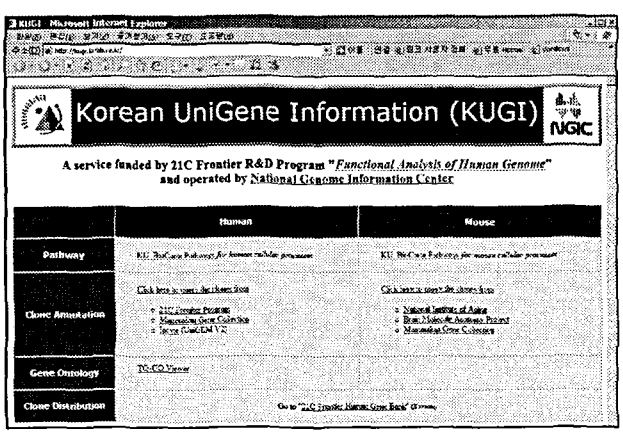

Fig. 1 Main Web Page of Korean UniGene Information (KUGI)

The KUGI search system provides both keyword search (Fig.2) and homology search using BLAST (Fig.3). In

keyword search, KUG accession number, GenBank accession number, gene symbol, gene title can be specified as query keywords. We support a list of multiple queries.

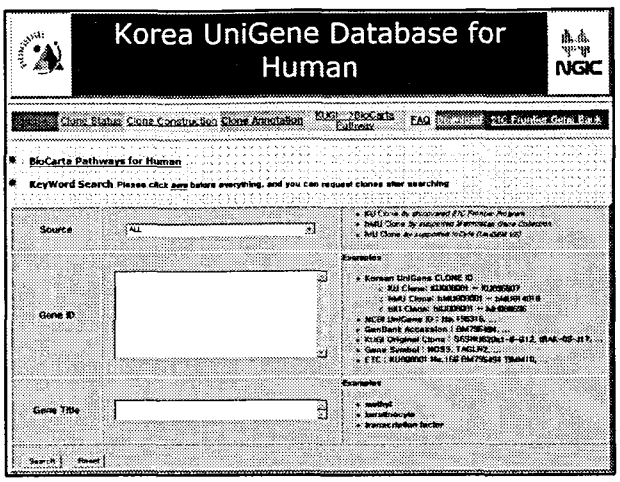

Fig. 2 Keyword Search Page

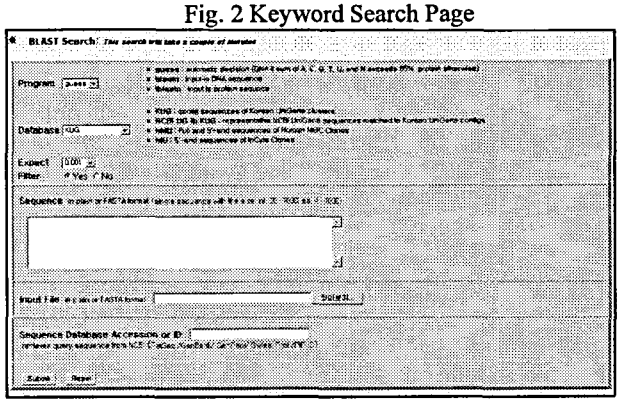

Fig. 3 Homology Search Page

Fig.4 shows query results. If one of the results is clicked, the detailed result is displayed as shown in Fig.5: 1) information of KUG clone and genome mapping, 2) annotation information, 3) external programs and database links, 4) sequences.

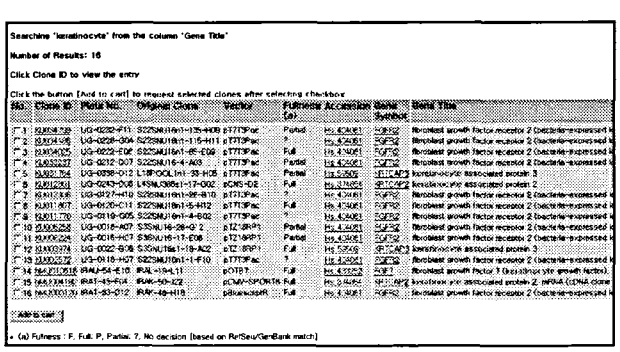

Fig. 4 Multiple results of query

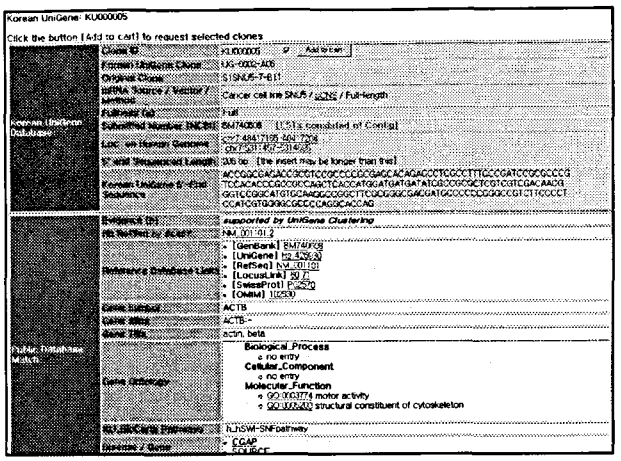

Fig. 5 One of multiple results

Fig.6 shows the mapping information of KUG clones on human genome by using UCSC genome browser. It also provides information for the genes analyzed with external databases such as ExPASY (http://www.expasy.org) and Ensembl (http://www.ensembl.org) databases.

The main page of the KU_BioCarta Pathways is shown in the Fig.7. Fig.8 displays the mapping information of KUG clones on BioCarta Pathway. With this, one can query either the genes on a certain pathway or the pathways a certain gene is involved.

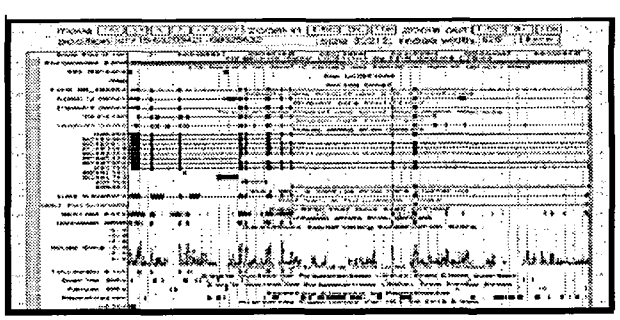

Fig. 6 Mapping KUG clones on Human Genome

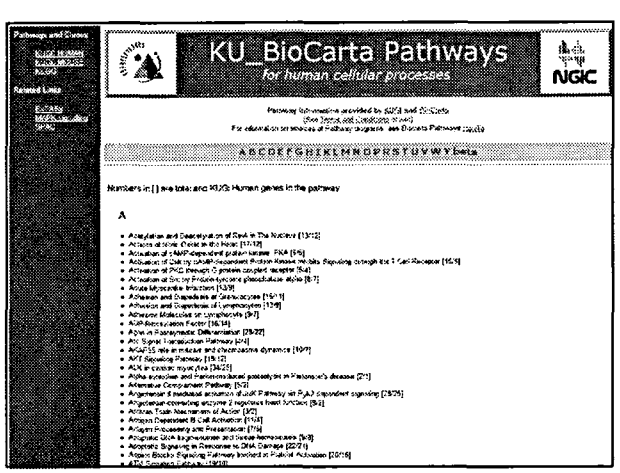

Fig. 7 Main page of KU_BioCarta Pathways

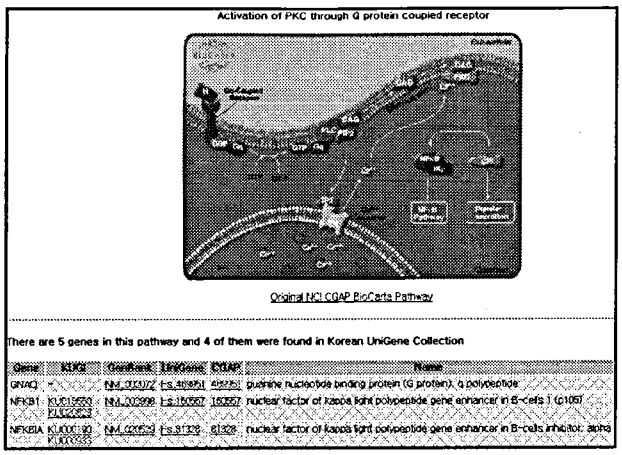

Fig. 8 Results of Mapping on KU_BioCarta Pathways

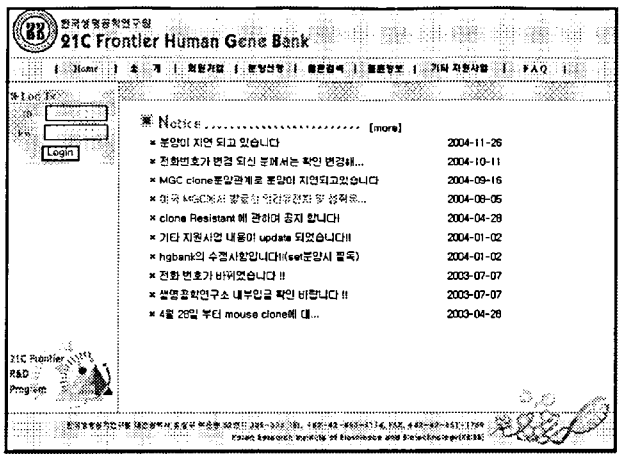

Fig. 10 Main Page of KUG Clone distribution

Homology search using BLAST was done in two ways: one with sequences and the other with GenBank accession number. If an accession number is given as a query, we first retrieve a sequence corresponding to the accession number, and then, run BLAST on the sequences retrieved. Fig.9 shows information about running parameters, the input sequence, databases to select, and various options. Each accession number in the query result is linked to external databases.

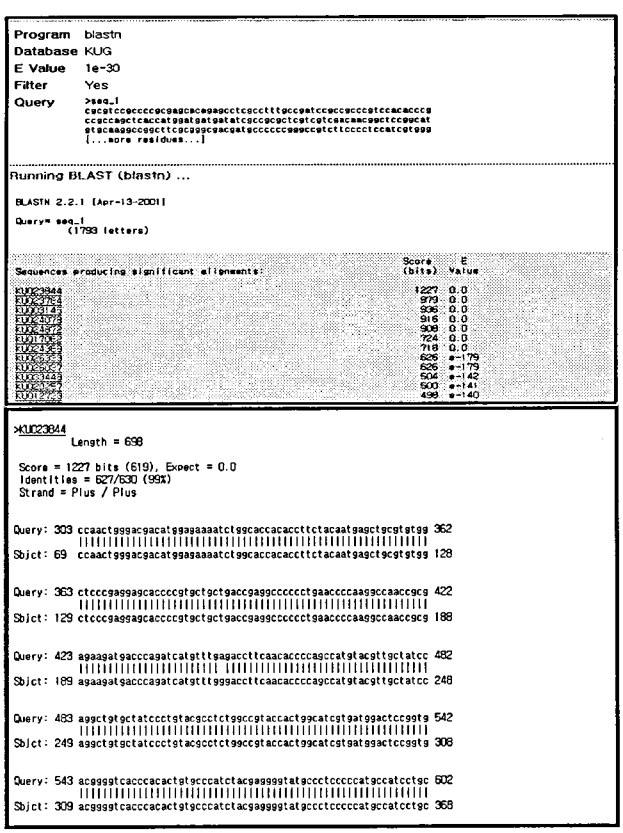

Fig. 9 Sequence Homology Results

3 CONCLUSION

We collected genes that are expressed in gastric, liver, brain, skin and so on cancer by large-scale sequencing in order to better understand the genetic events associated with these carcinogenesis[17]. The resulting high-quality ESTs were submitted to NCBI's dbEST database. Libraries discoveried by ours have been made available in the International Database of Cancer Gene Expression (http://cgap.nci.nih.gov). Especially, the 56% of the gastric ESTs in this database were represented by our ESTs. We have developed the KUGI (Korean UniGene Information) database and the query search system. We have introduced the process of collecting Korean UniGene clones, of building the database and pathway, and of building the web-based query system.

The KUGI database does not provide only information on full-length cDNAs and some novel genes but also distribution of gene ontology (GO), diseases, and pathways of clones. In the web-based search engine for the KUGI database, we provided two types of search interfaces: attribute-based search and similarity search of sequences. To support attribute search, we use DBMS technology; to support similarity search, we use BLAST that supports various homology of DNA sequences.

We believe that the KUGI database and search system provide very useful information that can be used to elucidate the causes and to develop the diagnostics and therapeutics of the cancers prevalent in Korean population. KUGI is available at http://kugi.kribb.re.kr/ and will continue to expand, incorporating our in-house data and others.

ACKNOWLEGEMENT

This work was supported by grant No. M103KB010023-04K0201-02310 of 21C Frontier Functional Human Genome Project From Ministry of Science & Technology of Korea. We appreciate all

researchers at The Center for Functional Analysis of Human Genome for their excellent technical support for construction of libraries and EST sequencing.

REFERENCES

[1] M. Adams et al. Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252, 1651-1656, 1991

[2] B Kim et al. Expression profiling and subtype-specific expression of stomach cancer. Cancer Res 63, 8248-8255, 2003

[3] D. Karan, DL Kelly, A Rizzino, MF Lin, SK Batra Expression profile of differentially-regulated genes during progression of androgen-independent growth in human prostate cancer cells. Carcinogenesis 23, 967-975, 2002

[4] J. M. Kim et al. Identification of gastric cancer-related genes using a cDNA microarray containing novel expressed sequence tags expressed in gastric cancer cells. Clin Cancer Res 11, 473-482, 2005

[5] Y. Suzuki et al. Statistical analysis of the 5' untranslated region of human mRNA using "Oligo-Capped" cDNA libraries. Genomics 64, 286-297, 2000

[6] M. A. van Driel et al. GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases. Nucleic Acids Res. 33, W758-W761, 2005

[7] J. H. Oh et al. Construction of multi-purpose vectors pCNS and pCNS-D2 are suitable for collection and functional study of large-scale cDNAs. Plasmid 51, 217-226, 2004

[8] M. B. Soares et al. Construction and characterization of a normalized cDNA library. Proc Natl Acad Sci 91, 9228-9232, 1994

[9] N. S. Kim et al. Gene cataloging and expression profiling in human gastric cancer cells by expressed sequence tags. Genomics 83, 1024-1045, 2004

[10] A. Hotz-Wagenblatt et al. ESTAnnotator: a tool for high throughput EST annotation. Nucleic Acids Res. 31, 3716 – 3719, 2003

[11] F. Meyer et al. GenDB—an open source genome annotation system for prokaryote genomes. Nucleic Acids Res. 31, 2187 – 2195, 2003

[12] LD Hillier et al. Generation and analysis of 280,000 human expressed sequence tags. Genome Res 6, 807-828, 1996

[13] X. J. Min et al. TargetIdentifier: a webserver for identifying full-length cDNAs from EST sequences. Nucleic Acids Res. 33, W669 - W672, 2005

[14] G. Dennis et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biology 4, R60, 2003

[15] E. Altermann and TR Klaenhammer PathwayVoyager: pathway mapping using the Kyoto Enclyclopedia of Genes and Genomes (KEGG) database. BMC Genomics 6, 60, 2005

[16] F. Mao et al. Pathway Mapping with Operon Information: An Integer-Programming Method. IEEE Computational Systems Bioinformatics Conference (CSB'04), 642-643, 2004

[17] S. Kato et al. Construction of a human full-length cDNA bank. Gene 150, 243-250, 1994