APPLICATION NOTE Genomics & Informatics Vol. 7(1) 38-40, March 2009

EST Knowledge Integrated Systems (EKIS): An Integrated Database of EST Information for Research Application

Dae-Won Kim[1]†, Tae-Sung Jung[1]†, Young-Sang Choi[3]†, Seong-Hyeuk Nam[1,2], Hyuk-Ryul Kwon[1], Dong-Wook Kim[3], Han-Suk Choi[3], Sang-Heang Choi[1] and Hong-Seog Park[1,2]*

[1]Genome Research Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon 305-806, Korea, [2]Department of Functional Genome, University of Science and Technology, Daejeon 305-333, Korea, [3]Department of Multimedia Engineering, Mokpo National University, Jeonnam 534-729, Korea

Abstract

The EST Knowledge Integrated System, EKIS (http://ekis.kribb.re.kr), was established as a part of Korea's Ministry of Education, Science and Technology initiative for genome sequencing and application research of the biological model organisms (GEAR) project. The goals of the EKIS are to collect EST information from GEAR projects and make an integrated database to provide transcriptomic and metabolomic information for biological scientists. The EKIS constitutes five independent categories and several retrieval systems in each category for incorporating massive EST data from high-throughput sequencing of 65 different species. Through the EKIS database, scientists can freely access information including BLAST functional annotation as well as Genechip and pathway information for KEGG. By integrating complex data into a framework of existing EST knowledge information, the EKIS provides new insights into specialized metabolic pathway information for an applied industrial material.

Keywords: biological system, data mining, expressed sequence tag, functional annotation, metabolic pathways

Introduction

The advent of high-throughput genomic sequencing technologies (Gupta, 2008; Jeong, *et al*., 2008) has created an enormous amount of information and resulted in an increasing number of publicly available genomes and experimental data. As a result of the current genomics revolution, genome sequence data from well-known model organisms is abundantly discovered and annotated (Bhak, *et al*., 2008; Hubbard, *et al*., 2002; Karolchik *et al*., 2008; Mailman, *et al*., 2007; Pruitt, *et al*., 2007). However, the circumstance is different for novel organisms coming from smaller expressed sequence tag (EST) sequencing projects. Many EST results are currently created by small laboratories without bioinformatics computing systems which are commonly available to genomic sequencing centers. Furthermore, these laboratories are generally lack of sufficient funds for sequencing, analyzing and managing the data. As such, a deluge of genomic data is difficult to control without appropriate funding for sequence characterization and genome data management for the purposes of effective access by biologists and the general public. To solve this problem, we have recently advanced a not-for-profit initiative for genomic sequencing and application research of biological model organism projects in order to provide extensive genome data to the scientists. We have also developed a user-friendly web site to facilitate access to sequences information of novel creatures.

Here, we demonstrate a newly developed website, EKIS, which stands for EST Knowledge Integration System. This website facilitates the annotation of new sequences with EC numbers and KEGG pathways based on BLASTX homology searches against the UniProt database using the PESTAS (http://pestas.kribb.re.kr/; unpublished). The EKIS is utilized not only as a generic website for obtaining genome project statistics, such as samples, libraries, analyzing ESTs, and annotated ESTs in current ongoing projects, but also as a mining tool for the KEGG pathway based on annotated EC numbers, expression profiling, homology searches and retrieval information systems.

Features and Results

Method

The EKIS, based on information derived from 65 EST sequencing projects (May 2008), is an integrated knowledge database developed using cutting-edge internet technology (Fig. 1a). To construct the data integrated

*Corresponding author: E-mail hspark@kribb.re.kr
Tel +82-42-879-8132, Fax +82-42-879-8139
Accepted 8 January 2009
†The authors wish it to be known that, in their opinion, the first three authors listed should be regarded as joint First Authors.

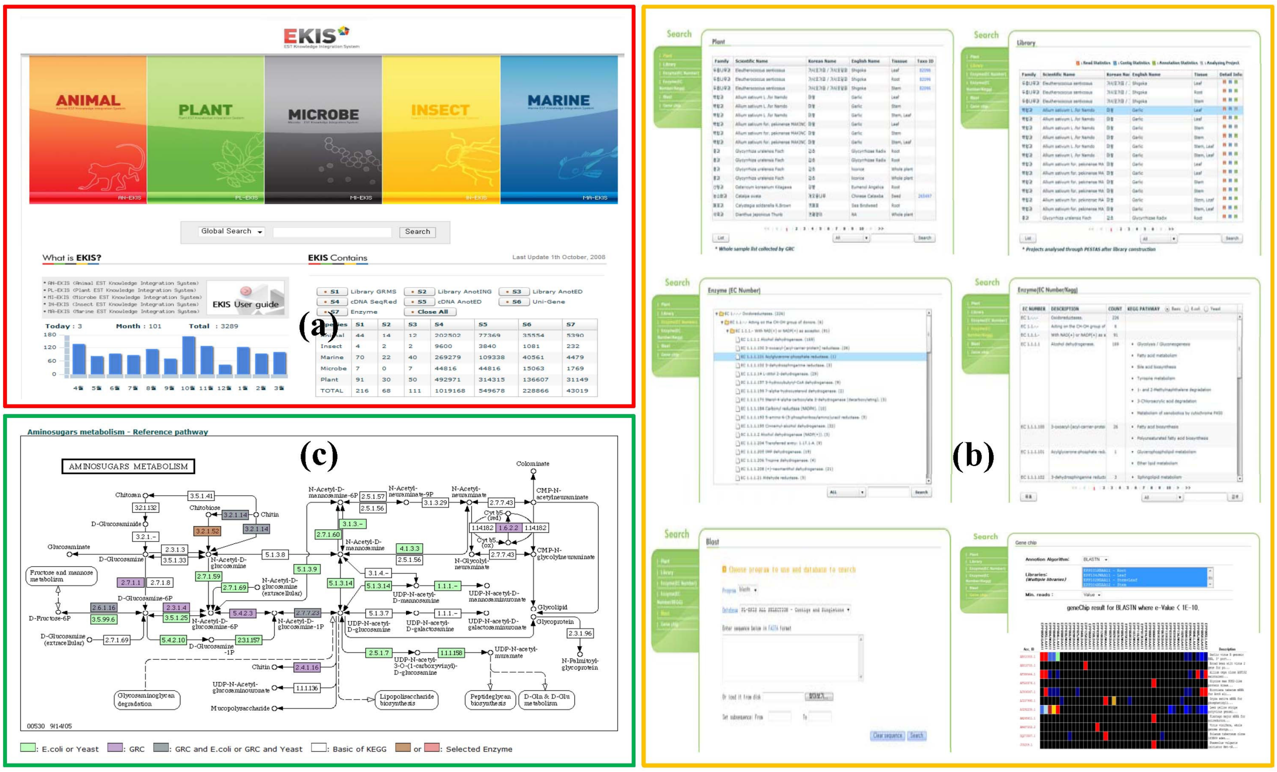

Fig. 1. Screenshot of EKIS graphical display. The interface contains three sections. The red box in the top left section shows the main page of EKIS. The orange box in the left section displays a variety of retrieval systems (ongoing EST projects, sample information, constructed library, enzyme and pathways, blast search and Genechip). The green box at the bottom left illustrates the KEGG map with a graphical representation using a variety of colors. See user guide for more details regarding EKIS.

system for EKIS, all sequenced EST data was annotated using the PESTAS (http://pestas.kribb.re.kr/pestas.jsp), mainly using BLASTX against public protein databases (NCBI NR database, UniProtKB-TrEMBL, UniProtKB-Swiss-Prot and KEGG), and these are fundamental information for establishing EKIS. The processed information are stored in relational tables designed in a typical key/value pair fashion as the physical using a MySQL database, and are indexed for retrieval and quick data access. The web interface was implemented in JavaScript and Adobe Flex running on an Apache web server.

Implementation

The EKIS is composed of five main categories (animal, plant, microbial, insect and marine) (Fig. 1a) and contains expressed transcript networks for 32,803 enzymes (7,465 for oxidoreductases, 9,616 for transfereases, 9,633 for hydrolases, 2,423 for lyases, 1,465 for isomerases, and 2,201 for ligases base on the enzyme commission). Each category consists mainly of six types of pages (Fig. 1a and 1c): a project outline, a construction of library, an EST projects summary, an integrated information retrieval, a pathway search, and a drug-compound page. The pathway search page is the core of the EKIS which is designed for retrieving pathway information for KEGG (Fig 1c). To construct these pathways, we extracted EC numbers from the description of UniProt results, and these EC numbers were mapped to a KEGG pathway map with graphical images. In addition, we have further developed the concept of Genechip to detect gene expression differences in each project by making a comparison with the count of accession id containing actual sequences from BLAST functional annotations.

Result

The EKIS contains an index of all annotation information taken from GEAR projects. With this integrated database, we are now able to find all pathways that belong to a specified EC number and enzyme name in each project of the source databases and obtain a list of the corresponding project or/and transcript IDs for cross-linking to each of the original databases. This provides EST project summary information, including the collected sample, library, analyzing EST and annotated EST project list (Fig. 1b). It should be noted that the EKIS does not include the entire contents of each source project database (e.g. raw EST data, all functional annotation information, detailed project information, etc.), but does provide the necessary fundamental project or/and transcript IDs through which users can access the original project databases to obtain all available information. If a user wants to access complete information about a certain project, he or she should contact the host of that project individually.

Future perspectives

In order to achieve a complete understanding of the biological pathways of an organism, it is essential to monitor the complete complement of all small molecule metabolites found in a specific tissue or organism at the metabolome level and gene expression at the transcriptome level, respectively. The important next step towards this goal will include the integration of experimental data on enzymes resulting from all fields of functional genomics as well as an analysis of the quantitative and qualitative collection of virtually all metabolites. In addition, this will require full-length cDNA data, which is a particularly important resource for determining their structural features. Ultimately, we will construct a discovery system for industrial enzymes in pursuit of a multi-purpose bioinformatics-cheminformatics-medical informatics system with a strong focus on new financial opportunities and novel intellectual properties.

Acknowledgements

This study was supported by grant M10752000001-07N5200-00110 from the Ministry of Education, Science and Technology and grant KGM1230812 from the Korea Research Institute of Bioscience and Biotechnology (KRIBB).

References

Bhak, J., Ghang, H., Reja, R., and Kim, S. (2008). Personal genomics, bioinformatics, and variomics. *Genomics & Informatics* 6, 161-165.

Gupta, P.K. (2008). Single-molecule DNA sequencing technologies for future genomics research. *Trends Biotechnol.* 26, 602-611.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Huminiecki, L., Kasprzyk, A., Lehvaslaiho, H., Lijnzaad, P., Melsopp, C., Mongin, E., Pettett, R., Pocock, M., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik, I., and Clamp, M. (2002). The ensembl genome database project. *Nucleic Acids Res.* 30, 38-41.

Jeong, H., and Kim, J.F. (2008). An optimized strategy for genome assembly of sanger/pyrosequencing hybrid data using available software. *Genomics & Informatics* 6, 87-90.

Karolchik, D., Kuhn, R.M., Baertsch, R., Barber, G.P., Clawson, H., Diekhans, M., Giardine, B., Harte, R.A., Hinrichs, A.S., Hsu, F., Kober, K.M., Miller, W., Pedersen, J.S., Pohl, A., Raney, B.J., Rhead, B., Rosenbloom, K.R., Smith, K.E., Stanke, M., Thakkapallayil, A., Trumbower, H., Wang, T., Zweig, A.S., Haussler, D., and Kent, W.J. (2008). The UCSC genome browser database: 2008 update. *Nucleic. Acids Res.* 36, D773-779.

Mailman, M.D., Feolo, M., Jin, Y., Kimura, M., Tryka, K., Bagoutdinov, R., Hao, L., Kiang, A., Paschall, J., Phan, L., Popova, N., Pretel, S., Ziyabari, L., Lee, M., Shao, Y., Wang, Z.Y., Sirotkin, K., Ward, M., Kholodov, M., Zbicz, K., Beck, J., Kimelman, M., Shevelev, S., Preuss, D., Yaschenko, E., Graeff, A., Ostell, J., and Sherry, S.T. (2007). The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.* 39, 1181-1186.

Pruitt, K.D., Tatusova, T., and Maglott, D.R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic. Acids Res.* 35, D61-65.