Genomics & Informatics Vol. 3(3) 94-97, September 2005

Patome: Database of Patented Bio-sequences

SeonKyu Kim[1] and ByungWook Lee[1,2]*

[1]National Genome Information Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon 305-333, Korea
[2]Department of BioSystems, Korea Advanced Institute of Science and Technology, Daejeon 305-701, Korea

Abstract

We have built a database server called Patome which contains the annotation information for patented bio-sequences from the Korean Intellectual Property Office (KIPO). The aims of the Patome are to annotate Korean patent bio-sequences and to provide information on patent relationship of public database entries. The patent sequences were annotated with Reference Sequence (RefSeq) or NCBI's *nr* database. The raw patent data and the annotated data were stored in the database. Annotation information can be used to determine whether a particular RefSeq ID or NCBI's *nr* ID is related to Korean patent. Patome infrastructure consists of three components-the database itself, a sequence data loader, and an online database query interface. The database can be queried using submission number, organism, title, applicant name, or accession number. Patome can be accessed at http://www.patome.net. The information will be updated every two months.

Keywords: Patome, Korean patent bio-sequences, annotation, web service.

Introduction

Recent advances in high-throughput sequencing technologies have enabled us to determine many genomic sequences fast and cheaply. The number of bases in GenBank has doubled approximately every 14 months (Benson *et al.*, 2004). These sequence data can be used in a variety of the industrial fields such as medicine, agriculture, nutrition, and environment (Collins *et al.*, 2003). For bio-sequences such as DNA, RNA, and protein, acquiring their intellectual property rights means to hold possession of comprehensive rights on their applications. One of the main issues in the post genomic era is how fast sequences and functions of genes are determined, and who preempts their patent rights.

Currently, three organizations offer patented bio-sequence data to the public: the National Center for Biotechnology Information (NCBI), the European Bioinformatics Institute (EBI), and the DNA Database of Japan (DDBJ). These organizations provide publicly patented bio-sequence data downloading, simple keyword searching, and alignment searching (Miyazaki *et al.*, 2004; Kanz *et al.*, 2005; Benson *et al.*, 2004). Thompson Derwent and the Chemical Abstract Service maintain commercial patent databases containing more comprehensive information. The commercial resources are expensive to use and are usually accessed by private commercial entities (Rouse *et al.*, 2005).

According to the KIPO report of May 2005, the office had received 6,237 bio-sequence related applications containing 117,873 nucleotide sequences and 57,926 protein sequences. On average, 100 bio-sequences are filed each month. We receive Korean patented bio-sequence data from KIPO once a month. The main purpose of filing patent application is acquiring legal rights which permit an inventor to prevent others from making, using or selling the invention, not finding their biological functions or meanings. So far, no attempt has been made to annotate patented bio-sequences from a biological perspective.

In this report, we describe Patome, a new publicly available database for annotated Korean patent sequences, as well as raw patent data from KIPO. To annotate the patented bio-sequences, BLAST algorithm (Altschul *et al.*, 1997) was used to compare Korean patented bio-sequences against two public databases: RefSeq or NCBI's *nr*. A database entry with the highest scoring BLAST hit was assigned to each patented bio-sequence. We also provide an online BLAST query interface to compare an input sequence against Korean patented bio-sequences. These data and services are available at http://www.patome.net/ and the information will be updated every two months.

*Corresponding author: E-mail bulee@kribb.re.kr,
Tel +82-42-879-8535, Fax +82-879-8519
Accepted 14 July 2005

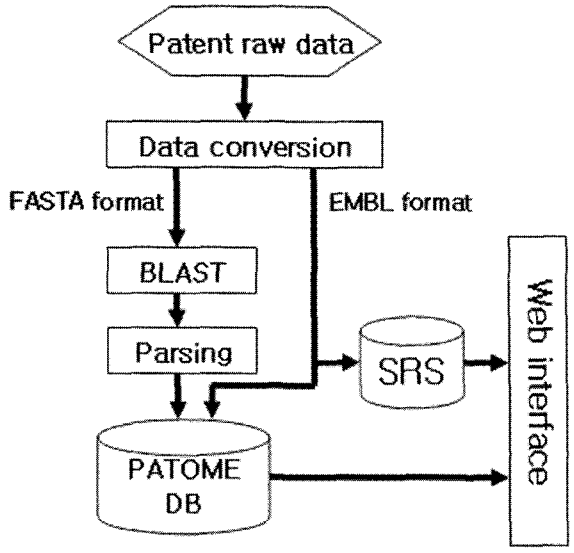

Fig. 1. The processing flow of patented data in the Patome

Methods

System Architecture

Patome infrastructure consists of three components-the database itself, a data loader, and an online database query interface (Fig. 1). It is operated on a Linux system. The consolidated dataset is stored and served from a MySQL relational database management system (http://www.mysql.com/). The data loader and online database query programs are implemented using JAVA (http://www.cpan.org/) standard library and Java server pages (JSP) (http://java.sun.com/products/jsp/), respectively. Web service is served on TOMCAT (http://jakarta.apache.org/tomcat/) framework ver. 5.1.

Data Acquisition

In accordance with the agreement between KIPO and KRIBB on July 8, 2004, KIPO provides NGIC with patented bio-sequence data in Korea once a month. The format of the data offered by KIPO was RDBMS dumps. Therefore, the data were converted to a FASTA format and an EMBL format for further processing.

Annotation of Patented Sequences

Patented bio-sequences can be divided into two types: *nucleic acid and amino acid. To annotate the nucleic acid* sequences, MegaBLAST (Zhang *et al.*, 2000) was used to compare the nucleic acid sequences against RefSeq mRNA. For those not covered by RefSeq mRNA, we annotated with NCBI's *nr* by using BLASTX. The BLAST results were filtered by an E-value cutoff score of 1.0E-5. Out of 117,873 patented nucleic acid sequences, 21,538

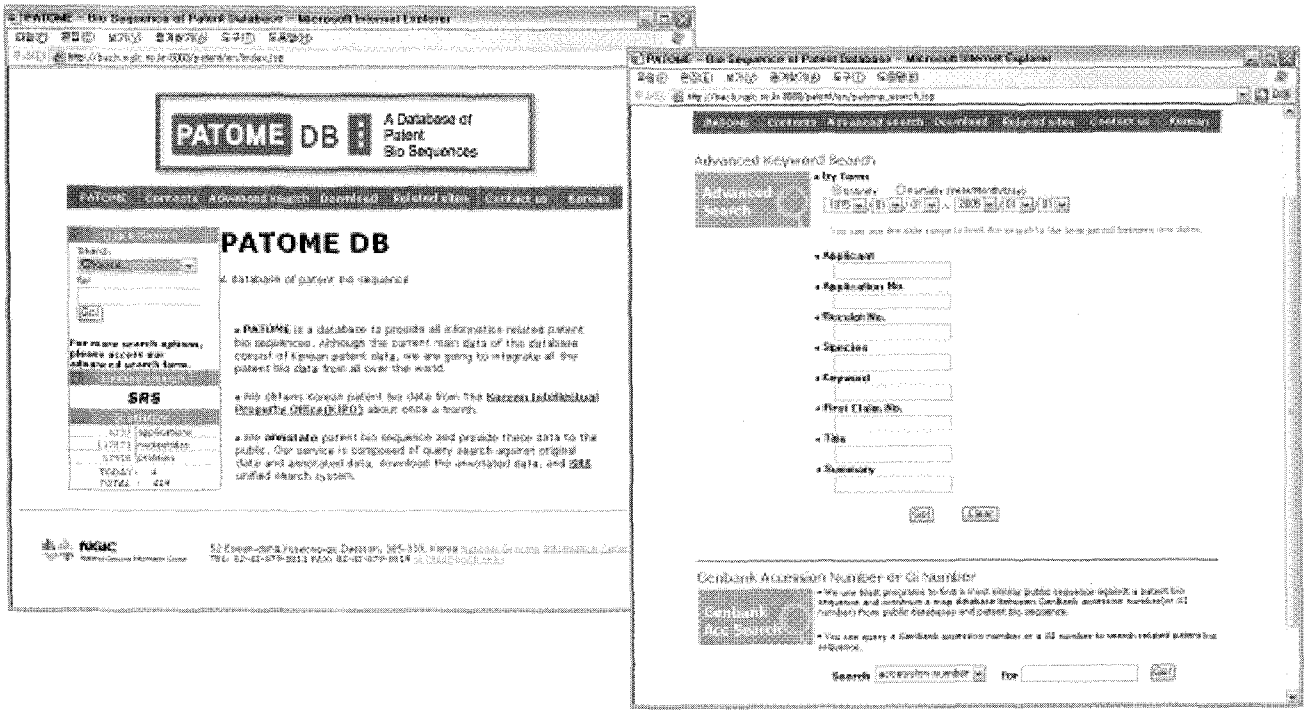

Fig. 2. Snapshots of Patome website: the left window is a homepage and the right window is an advanced search interface.

(18%) were annotated with RefSeq mRNA or NCBI's *nr*. Among un-annotated nucleic acid sequences, 59,318 (62%) were primer sequences that have explicit primer descriptions in the patent title or are less than 30bp in length.

To annotate the amino acid sequences, BLASTP was used to compare the amino acid sequences against RefSeq protein or NCBI's *nr*. An E-value cutoff was set to 1.0E-5. 27,270 (47%) out of 57,926 patented amino acid sequences were annotated with the public databases. 18,393 (60%) of the un-annotated amino acid sequences were very short amino acid sequences (<15 residues). The annotated data can be downloaded from the Patome website.

Construction of the Database

We designed and constructed the Patome database containing the raw patent data and the annotated bio-sequence data. In the Patome database, there are ten database tables which are classified into two sections: application related part and bio-sequence related part. The tables in the application related part contain basic patent information such as sequence information, applicants, abstracts, and priority rights. The tables in the bio-sequence related part contain sequence raw data such as references, feature, and sequences, as well as annotation data. We developed a data loader program that has the following functions: confirming data format, loading data, and exporting EMBL/FASTA files.

Searching

The Patome search web interfaces consist of three types: keyword, accession number, and BLAST similarity (Fig. 2). In the keyword search, there are two search modes: quick and advanced. If a user wants to query multiple keywords simultaneously, the advanced search is appropriate. In the advanced mode, the user can also limit the results by specifying a date range or by applying the Boolean operations AND, OR, and NOT. The quick search mode was designed for those not familiar with advanced syntax. In the quick mode, if a search field is not fixed, the target of search is all the fields. In both modes, the user can query keywords against the following fields: applicant name, application number, species, index word, priority rights number, English title, and summary.

In the accession number search, we provide the searching service on the annotated bio-sequences. When the user queries a accession number or a GI number (NCBI Seq. UID), the system displays patented bio-sequences and information related to an input number. From the search results, the user can obtain the information on whether their sequence is related to Korean patents.

To construct BLAST searching system, we installed WWW BLAST server (http://www.ncbi.nlm.nih.gov/BLAST/download.html) and formatted patent bio-sequences by NCBI's formatdb program. Users can compare their sequences against Korean patent bio-sequences.

Korean patent bio-sequences and related data have been incorporated into Sequence Retrieval System (SRS) (http://srs.ngic.re.kr/srs/) at NGIC. SRS is a scalable data integration platform that is used at over 300 commercial and academic sites. The user can utilize the SRS system to find the answers to complicated questions related to patented bio-sequences, for example, finding the protein motifs of a selected patent protein sequence

Results and Discussion

Through the Patome website, we offer the following services: the latest patented bio-sequences downloading, keyword search, similarity search, and tab delimited annotation data downloading. However, downloading of all the patented bio-sequences data is not allowable due to the agreement between KIPO and KRIBB. To download all the data, users must obtain permission from KIPO.

Acknowledgements

We thank Dr. YoungGyun Cho at KIPO for helpful discussion. This work was supported by the Korean Ministry of Science and Technology under grant number M1-0437-01-0002. This work was also supported by the KRIBB Research Initiative Program. We thank Maryana Bhak and Jong Bhak for editing the manuscript.

References

Altschul, S.F., Madden, T.L., Schäffer, 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.

Benson, D.A., Karche-Mizrachi, I., Limpman, D.J., Ostell, J., and Wheeler D.L. (2004). GenBank: Update. *Nucleic Acids Res.* 32, D23-D26.

Collins, F.S., Green, E.D., Guttmacher, A.E., and Guyer M.S. (2003). A vision for the future of genomics

research. *Nature* 422, 835-847

Kanz, C., Aldebert, P., Althorpe, N., Baker, W., Baldwin, A., Bates, K., Browne, P., van den Broek, A., Castro, M., Cochrane, G., *et al.* (2005). The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.* 33, D29-D33

Miyazaki, S., Sugawara, H., Ikeo, K., Gojobori, T., and Tateno, Y. (2004). DDBJ in the stream of various biological data. *Nucleic Acids Res.* 32, D31-D34.

Rouse, J.D., Castagnetto, J., and Niedner, R.H. (2005). PatGen-a consolidated resource for searching genetic patent sequences. *Bioinformatics* 21, 1707-1708.

Zhang, Z., Schwartz, S., Wagner, L., and Miller, W.(2000). A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.* 7, 203-214.