APPLICATION NOTES

Genomics & Informatics Vol. 5(2) 89-91, June 2007

RepWeb: A Web-Based Search Tool for Repeat-Related Literatures

Taeha Woo[1,2,¶], Younguk Kim[1,¶], Jekeun Kwon[1] and Jungmin Seo[1,2*]

[1]Korean BioInformation Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon 305-806, Korea
[2]Genome Research Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon 305-806, Korea

Abstract

Repetitive sequences such as SINE, LINE, and LTR elements form a major part of eukaryotic genomes. A literature search tool that summarizes the information contained within repeat elements would provide biologists in the field of genomics with a useful tool for analyzing genomic sequence features. We developed a java program designed to make literature access easier by using two search engines simultaneously. RepWeb is a web-based search system that provides a user friendly interface for searching the reference data and journals for information related to repeat elements by using the search engines, Google Scholar and PubMed, simultaneously. It provides an interface that displays the repeat element- related biological information, and includes useful functions such as the production of a repeat tree, clickable links to PubMed and Google Scholar, exporting, and sorting a field into date, author, journal and title.

Availability: RepWeb is freely available from the following web addresses: http://www.repeatome.org
http://www.repweb.org
http://bioportal.kobic.re.kr:8080/RepWeb

Keywords: repeat element, PubMed, Google Scholar, E-Utilities

Introduction

Repetitive DNA sequences include simple repeats and transposon-derived interspersed repeats such as SINE, LINE, and LTR transposons (Jurka *et al*., 2000). These sequences comprise a large part of the genomes of higher organisms. Despite the value of repetitive sequences, their role in many organisms is largely unknown, and the search tools which allow users to explore literature related to repeat terms are not well developed.

PubMed is one of the most important and powerful bibliographical information resources for biologists, and includes links to full text articles and other related resources (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=pubmed). However, PubMed is not optimized to search for specific interests such as repeat elements.

The search engine, Google Scholar, provides a simple way to search for scholarly articles on a broad range of topics, within a large database. The Google Scholar service is based primarily on the full-text archives of some of the largest scholarly publishers, combined with some open access A/I databases, preprint and reprint repositories, and pages from presumably academic WEB sites. It provides not only a link to the cited and citing articles, but also calculates and reports the citation scores of those articles (Jacso, 2005). Its purpose is to locate scholarly literature across all disciplines in many formats and to offer the best scholarly search experience for users (Dean, 2005).

Repbase is a database of prototypic sequences of repetitive DNA from different eukaryotic species. Repeat library information can be acquired from Repbase, which is one of several databases for repeat genomic elements. However, Repbase does not provide information pertaining to journal articles that contain repeat terminology.

The Entrez Programming Utilities (E-Utilities) are comprised of seven server-side programs that are enlisted to provide a stable interface for the Entrez query and database utility at the National Center for Biotechnology Information (NCBI). The E-Utilities are based on a preset URL syntax designed to translate a set of input parameters into appropriate values that can be used with NCBI software to search for and recover the requested information. The E-Utilities make up the ordered interface to the Entrez system, and include 23 databases covering diverse sets of biomedical data, which include gene records, nucleotide and protein sequences, three-dimensional molecular structures, and biomedical literature (http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=coursework.chapter.eutils).

In this paper we propose a web-based system, RepWeb, for searching reference information and journals related to repeat elements, which provides a user friendly interface. This is the first literature search web tool specifically designed to search for repeat elements reported to date.

¶Both of these authors contributed equally to the work.
*Corresponding author: E-mail jmseo@kribb.re.kr,
Tel +82-42-879-8133, Fax +82-42-879-8139
Accepted 3 April 2007

Features of RepWeb

For the terms and classification of repeats, text and repeat information were obtained from the RepeatMasker library of Repbase (Ver. 11.10, 11-13-2006) for each repeat element (Jurka *et al*., 2000) (Fig. 1). Repbase is a database of prototypic sequences representing repetitive DNA from different eukaryotic species. RepWeb relies on repeat dictionary, query terms, and word filters provided by Google scholar and E-Utilities (Fig. 2). We can provide various filters (article filters, author filters, subject area filters, etc.) for these queries using E-Utilities and Google Scholar. For example, if a keyword has *"Alu"* in the repeat name search box, its search terms will be as follows in PubMed : *"Alu" and (repeat [tw] or "repetitive element" [tw] or "repetitive sequences" [tw] or "Repetitive Sequences, Nucleic Acid" [MH] or "Interspersed Re petitive Sequences" [MH])*. In Google Scholar, they will be : *intitle:alu + (repeat OR "repetitive element" OR "repetitive sequences" or "Repetitive Sequences, Nucleic Acid" OR "Interspersed Repetitive Sequences")*.

These results are automatically parsed and assigned to each search engine when submitted. RepWeb also has a function to allow users to download each search result text file and its link to the Repbase database in the GIRI. The Web crawler and parser are implemented by Python, and Google Scholar is used to acquire more information from the WEB. The user can search using a specific search

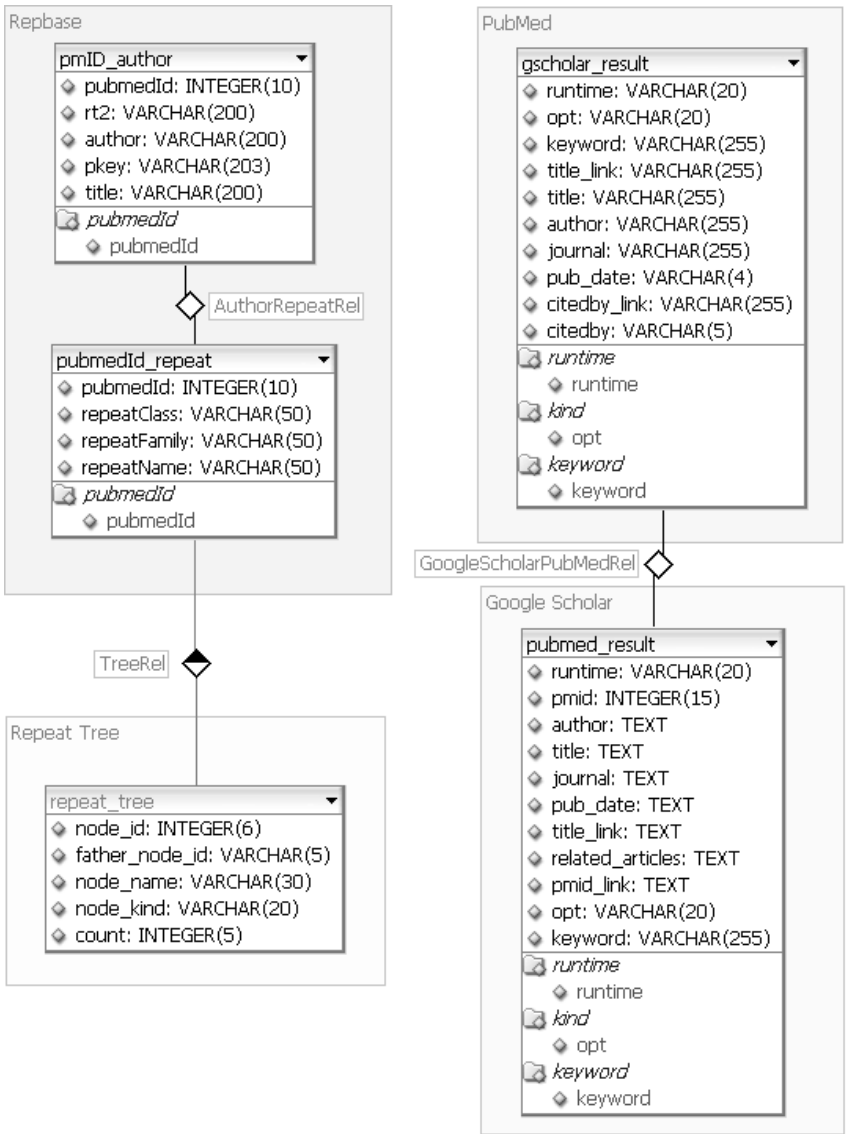

Fig. 1. Database Schema in RepWeb.

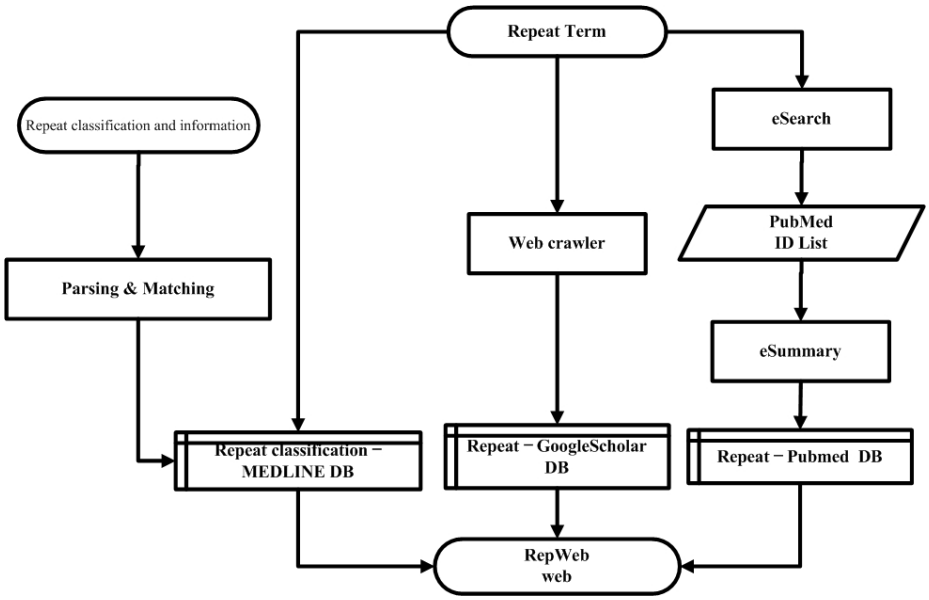

Fig. 2. Flowchart for searching a repeat term in RepWeb.

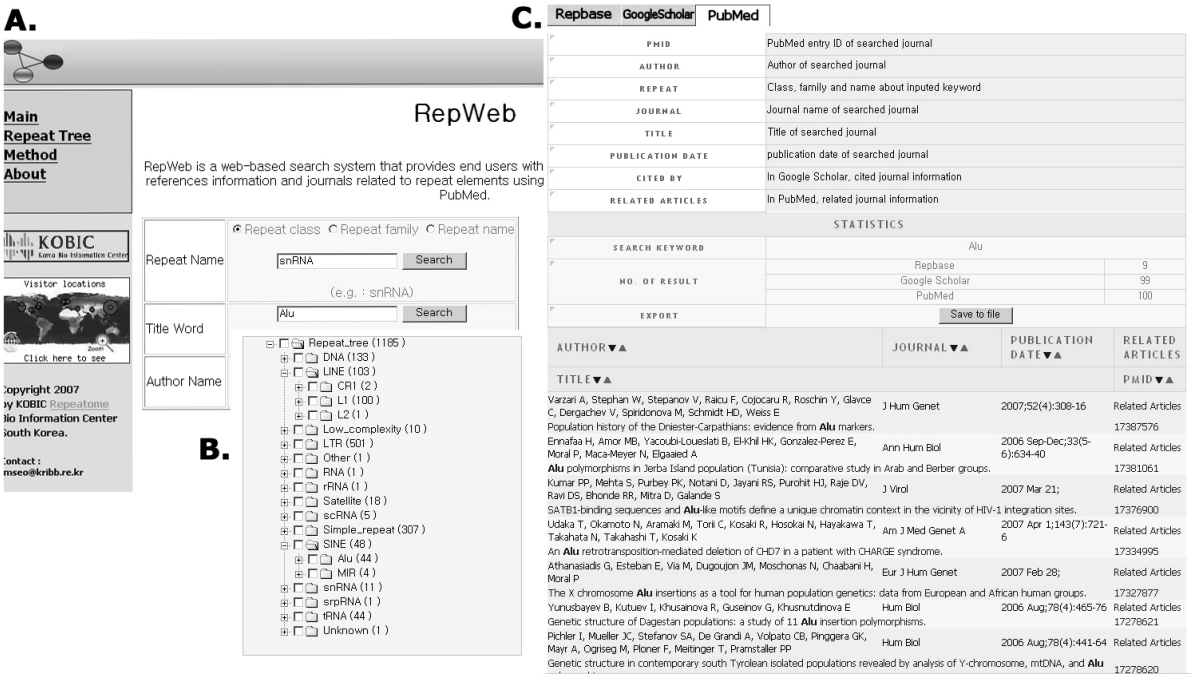

Fig. 3. Web interface of RepWeb. (A) Search interface of RepWeb. The user is able to search the repeat term with four inputs. (B) Repeat Tree. This tree is classified according to repeat class, family and name. In addition, users can search repeat information for the selected repeat. (C) Result of searching in Repbase with a table view.

term which will return hundreds of hits. After the initial search, Web Crawler delivers the results and parses the information. The final results are presented as a table on the web (Fig. 3C).

The web server, RepWeb, has been developed to search journals related to the repeats in Repbase, Google Scholar and PubMed. Through the web interface, researchers are able to retrieve journals pertaining to the queried keyword within three resources. To acquire the journal-related Repbase, RepWeb constructs a local

relational database which integrates two public databases; repeat library data and PubMed. RepWeb is a web interface based on a Tomcat server with dynamic content generated by a Java Server Page (JSP) and AJAX (Asynchronous JavaScript and XML). By applying these technologies, RepWeb provides users with a dynamic, responsive, and faster browser. The interface is presented as a repeat tree menu that allows users to search easily while selecting a particular repeat (Fig. 2B).

Results

RepWeb is a Java application that provides a user-friendly interface for searching the reference information and journals related to repetitive sequences within the genomes of various organisms. RepWeb has a distinctive advantage, in that it retrieves the results from two search engine sites simultaneously. It relies on two databases, using the search term results from Google Scholar and PubMed, and has a Repbase database which matches the Repbase library with information from PubMed. The application is freely available from http://www.repeatome.org or http://www.repweb.org or directly from http://bioportal.kobic.re.kr:8080/RepWeb.

Future Directions

We will update the search function of RepWeb by sequence, and apply it to a gene pathway using our pipeline and web database system. RepWeb will continue to incorporate newly established repeat elements. The role of RepWeb in providing literature information for biologists will continue to adapt in line with new developments and the needs of researchers in this field.

Acknowledgements

We are grateful to anonymous RepWeb reviewers for their comments. This work was supported by the Korean Ministry of Science and technology under grant number M10508040002-06N0804-00210 and M10407010001-06N0701-00110. This work was also supported by the KRIBB Research Initiative Program.

References

Dean, G. (2005). A look at Google Scholar, PubMed, and Scirus: comparisons and recommendations., In *the Journal of the Canadian Health Libraries Association*, pp. 85-89.

Entrez Utilities. *http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html.*

Google Scholar. *http://schoar.google.com/scholar/about.html.*

Jacso, P. (2005). Comparison and analysis of the citedness scores in web of science And Google Scholar., In *Lecture Notes in Computer Science*, Fox, E.A., Neuhold, E.J., Premsmit, P., Wuwongse, V. ed. (Springer-Verlag, Berlin), pp. 360-369.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., Walichiewicz, J. (2000). Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet*. 16, 418-420.

PubMed. *http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=pubMed.*

RepBase homepage.Genetic Information Research Institute. *http://www.girinst.org/*