MedlineR: an open source library in R for Medline literature data mining

Simon M. Lin1,2,*, Patrick McConnell1, Kimberly F. Johnson1 and Jennifer Shoemaker1,2

1Duke Comprehensive Cancer Center and 2Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, NC 27710, USA

Received on May 11, 2004; revised on June 21, 2004; accepted on July 4, 2004
Advance Access publication July 29, 2004

ABSTRACT
Summary: We describe an open source library written in the R programming language for Medline literature data mining. This MedlineR library includes programs to query Medline through the NCBI PubMed database; to construct the co-occurrence matrix; and to visualize the network topology of query terms. The open source nature of this library allows users to extend it freely in the statistical programming language of R. To demonstrate its utility, we have built an application to analyze term-association by using only 10 lines of code. We provide MedlineR as a library foundation for bioinformaticians and statisticians to build more sophisticated literature data mining applications.

Availability: The library is available from http://dbsr.duke.edu/pub/MedlineR.
Contact: Lin00025@mc.duke.edu

MOTIVATION
Currently, there is a substantial interest in linking microarray and proteomics data to literature (Masys, 2001). This literature data-mining step offers a sensible explanation of the gene list generated from high-throughput experiments by showing their associations. Many programs have been developed to facilitate this process by querying the Medline literature database, e.g. MedMiner (Tanabe et al., 1999), PubGene (Jenssen et al., 2001), XplorMed (Perez-Iratxeta et al., 2003) and PubMatrix (Becker et al., 2003). However, none of the currently available literature data-mining programs can be invoked within a statistical analysis program and the results cannot be easily incorporated for further statistical inference (Table 1). In the paper, we fill in this gap by providing an open source interface to Medline for literature data mining in a statistical programming language.

IMPLEMENTATION AND FEATURES
We implemented the MedlineR library in the programming language of R because R has extensive support of statistical modeling libraries. Moreover, R has already been used as a foundation to build bioinformatics libraries for microarray and genomic data analysis (www.bioconductor.org). With this MedlineR library, it is possible to streamline the microarray analysis from raw data processing, statistical gene list identification and finally to literature analysis. The source code of MedlineR is quality-controlled by a variation of pair programming (Williams and Kessler, 2002). A list of programs and example codes are summarized in Table 2. It includes the following two interfaces: querying Medline through the NCBI Pubmed HTTP service (Wheeler et al., 2002), and exporting data in the Pajek format. Pajek is a visualization software to examine the network topology interactively (Batagelj and Mrvar, 2004). The primary interface to Pubmed is by keyword-based retrieval that can be used together with Boolean operators. Common keywords can be mapped to the controlled vocabulary of MeSH (Lipscomb, 2000), which is used to index Medline abstracts. MedlineR itself can run on UNIX, LINUX, Windows and Mac operating systems where

*To whom correspondence should be addressed.
Table 2. List of procedures in the MedlineR library

<table>
<thead>
<tr>
<th>Program</th>
<th>Description</th>
<th>Example usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>findAssociation</td>
<td>An example application to mine the literature for association</td>
<td>findAssociation(termList=c(&quot;STE12&quot;, &quot;DIG1&quot;, &quot;HOG1&quot;, &quot;SST2&quot;))</td>
</tr>
<tr>
<td>queryPubmed</td>
<td>Query NCBI Pubmed</td>
<td>queryPubmed(&quot;term=cancer+OR+diabetes&quot;)</td>
</tr>
<tr>
<td>fetchAnAbstract</td>
<td>Retrieve a MEDLINE abstract</td>
<td>fetchAnAbstract(pmid=134567)</td>
</tr>
<tr>
<td>mapToMeSH</td>
<td>Map a term to the MeSH controlled vocabulary thesaurus</td>
<td>mapToMesh(term=&quot;nosebleed&quot;)</td>
</tr>
<tr>
<td>stopwords</td>
<td>Return a list of stop words, which are non-information-bearing words</td>
<td>stopwords()</td>
</tr>
<tr>
<td>highlightKeywords</td>
<td>Highlight keywords in a sentence</td>
<td>highlightKeywords(sentence=&quot;p53 downregulates VEGF&quot;, keywords=&quot;p53&quot;)</td>
</tr>
<tr>
<td>countApair</td>
<td>Count the number of abstracts with both terms</td>
<td>countApair(term1=&quot;STE18&quot;, term2=&quot;STE12&quot;)</td>
</tr>
<tr>
<td>getAmatrix</td>
<td>Return a co-occurrence matrix for a list of terms</td>
<td>getAmatrix(termList=c(&quot;STE12&quot;, &quot;DIG1&quot;, &quot;HOG1&quot;))</td>
</tr>
<tr>
<td>pauseBetweenQueries</td>
<td>Pause between queries, according to the NCBI rule</td>
<td>pauseBetweenQueries()</td>
</tr>
<tr>
<td>write.pajek</td>
<td>Output in Pajek format for visualization</td>
<td>write.pajek(result.matrix, fileName=&quot;genes.net&quot;)</td>
</tr>
</tbody>
</table>

Fig. 1. Schematic representation of an application built upon the MedlineR library. (A) Pairwise queries to PubMed to find associations in a list of terms. (B) Visualization of the pairwise data association matrix. (C) Network visualization of the association. (D) Code in R, using the MedlineR library.
R can be compiled, whereas Pajek only runs on Windows platform.

**AN EXAMPLE APPLICATION**

To demonstrate the utility of the MedlineR library, we have built an application to find the associations among a list of genes (Fig. 1). The result is visualized as a network: each node represents a gene and each edge represents a connection in the literature. Figure 1B is a static graph generated in R by the ‘image’ command. Figure 1C is an interactive display in the Pajek viewer. For a detailed discussion of associative literature mining, we refer readers to the paper of Sluka (2002).

**CONCLUSIONS AND FUTURE DIRECTIONS**

We provide MedlineR as an open source library for bioinformaticians and statisticians to build more sophisticated literature data-mining applications. The source code of MedlineR is available from http://dbsr.duke.edu/pub/MedlineR. To facilitate open development, feature request, bug tracking, code contribution and version control tools are provided at this website.

Currently, most of the medical literature is only available in the abstract format via Medline. With the increasing collection of full texts available at PubMedCentral (Eisen et al., 2002) and BMC OpenAccess (Owens, 2003), we are expecting to have new functions in the MedlineR library to handle full text.

**REFERENCES**


