Bioinformatics, 31(10), 2015, 1689–1691
doi: 10.1093/bioinformatics/btv016
Advance Access Publication Date: 12 January 2015
Applications Note

OXFORD

Data and text mining

QTLMiner: QTL database curation by mining tables in literature

Jing Peng[1,3,†], Xinyi Shi[3,†], Yiming Sun[2], Dongye Li[1], Baohui Liu[3], Fanjiang Kong[3] and Xiaohui Yuan[3,*]

[1]College of Electronic and Information, Northeast Agricultural University, Harbin, China, [2]School of Computer Science and Technology, Changchun University of Science and Technology, Changchun, China and [3]The Key Lab of Soybean Molecular Design Breeding, Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, Harbin, China

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Jonathan Wren

Received on August 29, 2014; revised on December 25, 2014; accepted on January 7, 2015

Abstract

Motivation: Figures and tables in biomedical literature record vast amounts of important experiment results. In scientific papers, for example, quantitative trait locus (QTL) information is usually presented in tables. However, most of the popular text-mining methods focus on extracting knowledge from unstructured free text. As far as we know, there are no published works on mining tables in biomedical literature. In this article, we propose a method to extract QTL information from tables and plain text found in literature. Heterogeneous and complex tables were converted into a structured database, combined with information extracted from plain text. Our method could greatly reduce labor burdens involved with database curation.

Results: We applied our method on a soybean QTL database curation, from which 2278 records were extracted from 228 papers with a precision rate of 96.9% and a recall rate of 83.3%, *F* value for the method is 89.6%.

Availability and implementation: QTLMiner is available at www.soyomics.com/qtlminer/.

Contact: yuanxh@iga.ac.cn

1 Introduction

A quantitative trait locus (QTL) is a region of DNA that is associated with a particular phenotypic trait, and is important information for breeding and gene cloning. QTL studies have a long and rich history; a large number of articles recording the experiment results of QTL mapping have been published in recent years. To allow researchers to query QTL information easily, several databases have been developed by manually curating the information scattered in the literature (Grant *et al.* 2010; Monaco *et al.* 2014). However, manual curation is labor-intensive work, and it is impossible to remain current with the rapidly increasing numbers of articles; consequently, automatic methods are necessary for QTL database development. Unlike other experiment results, such as genome and molecular data, QTL information is usually presented in tables, especially in scientific papers. It is difficult to extract such information using traditional text-mining methods that focus on mining knowledge in unstructured plain text. Although, knowledge hidden in tables and figures has received increasing attention (Thomas *et al.* 2010), as far as we know there are only a few works on web table mining (Wang *et al.* 2012; Yang *et al.* 2002) and no published work on mining tables in biomedical literature.

In this article, we present a pipeline for automatically mining QTL information from scientific literature. A flowchart of this process is shown in Figure 1. We started the process by locating QTL tables in articles; the tables were then converted into two-dimensional QTL databases. The Natural Language Process method was used to extract information that was not presented in tables, including captions, footnotes and full text. Lastly, content

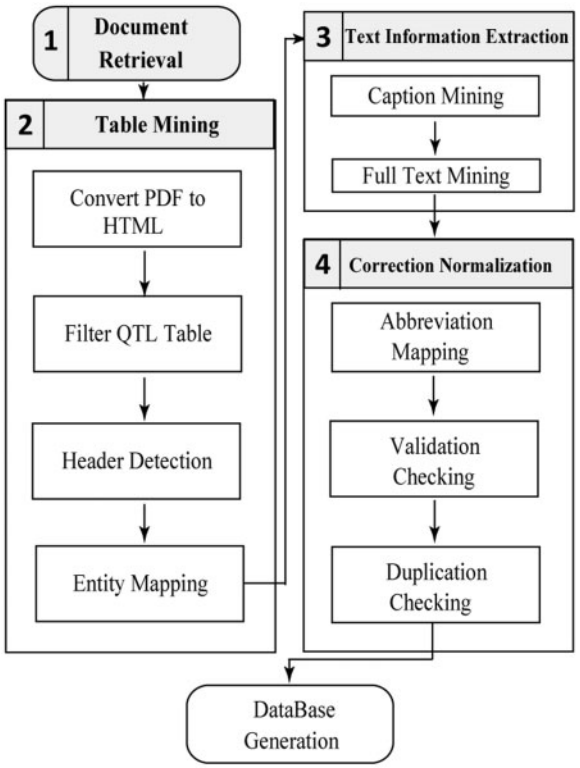

Fig. 1. Flowchart of QTL information mining

verification and duplication checks were performed on the outputs of previous steps. We applied this pipeline for soybean QTL database curation with a precision rate of 96.9% and *F* value of 89.6%; this software can also be used to construct QTL databases for other species.

2 Methods

2.1 Document retrieval

Document retrieval is always the first step in database curation. Curators rely on search engines such as PubMed and Google to filter related articles. However, results from search engines contain noise and greatly decrease text-mining performance. We manually selected articles containing soybean QTL information from the results returned by search engines.

2.2 Table mining

Scientific articles are usually distributed in PDF format and are not suitable for text-mining tasks, so we first converted the PDF files into HTML format. Contents between the HTML tags <Table> and <\Table> were then selected to reconstruct a table. Each table was stored in an Excel file so the location of the cells in each table could be defined by rows and columns. Table captions and legends were extracted and stored separately as plain-text files.

One QTL record has at least has two contents: a marker and a trait. Traits may be described in table captions or in other portions of a paper; however, markers are usually listed in tables. Consequently, we defined a table as a candidate QTL table if it recorded marker information.

Once a candidate QTL table was selected, our first step was to perform header identification. In this step, table headers were mapped to the 19 predefined fields of the soybean QTL database. For a regular table in which headers lay in the first row and no cells spanned multiple columns or rows, values in the first row were directed and mapped to the predefined database fields. For a complex table in which headers spanned multiple rows and columns, header identification became challenging. For example, complex headers may indicate that the table has more than one value associated with a field of one record. Such a table can be looked at as a fusion of several simple tables, so we divided one record in the table into several separate records. There were also many other cases and because of this we made 5 rules based on 128 articles containing soybean QTL information. Those rules cover 94% QTL tables.

Entities in the table were mapped to corresponding predefined database fields according to the columns in which they were located. Values of entities were checked to determine if they were in the range of fields.

2.3 Text information extraction

One table does not contain all the QTL information for one record, since information such as trait and population types are located in the caption, the legend, or other areas of the paper. Three kinds of methods were used to extract information from plain text. First, we scanned the table's caption and legend using a vocabulary dictionary, selecting sentences containing words describing plant traits and analysis methods. A simple template analysis was then applied to these sentences to extract information expressed in frequently used syntax patterns. Advanced templates based on a deep dependency tree (Richard *et al.* 2013) were used to find information that was not expressed in regular syntax patterns.

2.4 Correction and normalization

There are inevitably some mistakes and duplicate records caused by the file-format conversion or text-mining process. Corrections were performed on the table and text-mining results from three areas. (i) Abbreviation mapping: authors used some abbreviations in tables or full texts; we scanned the entire paper to find the full names associate with the abbreviations. (ii) Validation check: records with empty trait values or invalid markers were deleted from the final results. We also checked the validation of records by prior knowledge, such as correspondence between markers and linkage groups. (iii) Duplication check: records with the same values in the marker, trait, year, location, parents and method were considered the same records.

3 Results

We applied this pipeline for soybean QTL database construction, from which 2278 records were automatically extracted from 228 papers. The precision rate of the mining results was 96.9%. Compared with the records in the SoyBase QTL database (Grant *et al.* 2010), we collected more records from the literature. We also tested our pipeline on the literature listed on the SoyBase website, which was used for manual curation, with a recall rate up to 79.7%. This method is also suitable for development of comprehensive QTL databases for a wide range of species.

Funding

This research was funded by 'Hundred Talents Program' of the Chinese Academy of Sciences and Postdoctoral Science Foundation of Heilongjiang Province (LBH-Z13018).

Conflict of Interest: none declared.

References

Grant,D. *et al*. (2010) SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Res.*, **38**, D843–D846.

Monaco,M.K. *et al*. (2014) Gramene 2013: comparative plant genomics resources. *Nucleic Acids Res.*, **42**, D1193–D1199.

Richard,S. *et al*. (2013). Parsing with compositional vector grammars. In: *Proceedings of ACL 2013*.

Thomas,A. *et al*. (2010) Highlights of the BioTM 2010 workshop on advances in bio text mining, BMC Bioinf., **11**, I1.

Wang,J. *et al*. (2012) Understanding tables on the web. *LNCS*, **7532**, 141–155.

Yang,Y. and Luk,W.-S. (2002) A framework for web table mining. In: *Proceedings of the International Workshop on Web Information and Data Management*, pp. 36–42.