Gene 518 (2013) 132–138

Contents lists available at SciVerse ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Rule extraction in gene–disease relationship discovery

Wen-Juan Hou *, Hsiao-Yuan Chen

Department of Computer Science and Information Engineering, National Taiwan Normal University, Taipei, Taiwan

ARTICLE INFO

Available online 9 December 2012

Keywords:
Rule learning
Gene–disease association
Biomedicine
Text mining

ABSTRACT

Background: Biomedical data available to researchers and clinicians have increased dramatically over the past years because of the exponential growth of knowledge in medical biology. It is difficult for curators to go through all of the unstructured documents so as to curate the information to the database. Associating genes with diseases is important because it is a fundamental challenge in human health with applications to understanding disease properties and developing new techniques for prevention, diagnosis and therapy.
Methods: Our study uses the automatic rule-learning approach to gene–disease relationship extraction. We first prepare the experimental corpus from MEDLINE and OMIM. A parser is applied to produce some grammatical information. We then learn all possible rules that discriminate relevant from irrelevant sentences. After that, we compute the scores of the learned rules in order to select rules of interest. As a result, a set of rules is generated.
Results: We produce the learned rules automatically from the 1000 positive and 1000 negative sentences. The test set includes 400 sentences composed of 200 positives and 200 negatives. Precision, recall and *F*-score served as our evaluation metrics. The results reveal that the maximal precision rate is 77.8% and the maximal recall rate is 63.5%. The maximal *F*-score is 66.9% where the precision rate is 70.6% and the recall rate is 63.5%.
Conclusions: We employ the rule-learning approach to extract gene–disease relationships. Our main contributions are to build rules automatically and to support a more complete set of rules than a manually generated one. The experiments show exhilarating results and some improving efforts will be made in the future.

1. Introduction

Biomedical data available to researchers and clinicians have increased dramatically over the past years because of the exponential growth of knowledge in medical biology. One of the challenges that scientists in this domain face is that most of the relevant information remains hidden in the unstructured text of the published papers. It is difficult for curators to go through all of the documents so as to curate the information to the database. Thus, it is necessary to extract biomedical information automatically spread over several different databases. Associating genes with diseases is an important area of researches because it is a fundamental challenge in human health with applications to understanding disease properties and developing new techniques for prevention, diagnosis and therapy.

To date, a variety of methods has been published for exploring the relationships between genes and diseases. Some previous studies use protein–protein interactions to predict gene–disease relationships (Gottlieb et al., 2011; Hwang et al., 2011; Navlakha and Kingsford, 2010; Ozgur et al., 2008). Some approaches use gene ontology (GO) terms (Ashburner et al., 2000) or disease ontology (DO) terms (Kibbe et al., 2006) to compare the similarity between genes and diseases (Gaulton et al., 2007; Marthur and Dinakarpandian, 2010; Schlicker et al., 2010; Yilmaz et al., 2009). The genes that are associated with a particular disease are ranked based on some comparisons involving GO or DO terms. Other controlled vocabularies such as MeSH (Mottaz et al., 2008) have already been utilized for linking proteins to disease terminologies, MeSH. Gene expressions (Gaulton et al., 2007; Ma et al., 2007), protein sequences (George et al., 2006) and positional information (Maver and Peterlin, 2011) are also served as the important evidences to relating genes and diseases. Moreover, some researchers utilize the text mining techniques to extract gene–disease associations from the biomedical literature (Hou et al., 2011; Hristovski et al., 2005; Hwang et al., 2011; Yu et al., 2008). These works demonstrate that associating genes with diseases is an active area of researches as it can lead to better understanding of diseases and it can reduce time and expenditure in developing effective drugs and treatment.

In the paper, we use two resources for our experiments: Online Mendelian Inheritance in Man (OMIM) (Hamosh et al., 2002) and MEDLINE. OMIM is one of the most well-known databases that

Abbreviations: GO, Gene ontology; DO, Disease ontology; OMIM, Online Mendelian inheritance in man; NLM, National Library of Medicine; IR, Information retrieval; NLP-based, Natural language processing-based; MBSP, Memory-based shallow parser; ILP, Inductive logic programming; MLL, Myrloid–lymphoid leukaemia; MLL, Mixed-lineage leukaemia.

* Corresponding author at: Department of Computer Science and Information Engineering, National Taiwan Normal University, No.88, Sec. 4, Tingzhou Road, Wenshan Distinct, Taipei 116, Taiwan. Tel.: +886 2 7734 6660; fax: +886 2 2932 2378.
E-mail address: emilyhou@csie.ntnu.edu.tw (W.-J. Hou).

0378-1119/$ – see front matter.
http://dx.doi.org/10.1016/j.gene.2012.11.060

contains gene–disease annotations. It is curated by the NCBI and Johns Hopkins University. The full-text overviews in OMIM contain information on all known Mendelian disorders and over 12,000 genes. It is a comprehensive knowledge base of human genes and genetic diseases. For biomedical researchers and clinicians, OMIM serves as an important resource to support Mendelian inheritance information. One of the most comprehensive textual sources of biomedical information is MEDLINE. It contains more than 19 million citations from more than 7300 different publications dating from 1966, and it continues updating weekly by U.S. National Library of Medicine (NLM). Biological researchers often access MEDLINE abstracts or free full-text articles through the PubMed database (http://www.ncbi.nlm.nih.gov/pubmed/) or the information retrieval (IR) system. These systems return documents satisfying users' information needs. It is valuable for the biologists to mine information from the returned documents.

We aim at the methods of helping with relationship extraction to genes and diseases in this study. As mentioned before, many researches have been published for exploring the relationships between genes and diseases. If we do not restrict the targets on genes and diseases, more approaches to extracting relations of interest have been reported in the literature. The first one is a manually generated template-based/rule-based method which uses patterns generated by domain experts to extract concepts connected by a specific relation from text (Feldman et al., 2002; Fundel et al., 2007; Yu et al., 2002). The second one is an automatic template/rule-learning method which creates similar templates/rules automatically by generalizing patterns/rules from text surrounding concept pairs known to have the relationship of interest (Gopalakrishnan et al., 2010; Kim et al., 2007; Tatar and Cicekli, 2009). The third one is a statistical method which identifies relationships by looking for concepts that are found with each other more often than would be predicted by chance (Lindsay and Gordon, 1999). Finally, Natural Language Processing-based (NLP-based) methods perform a substantial amount of sentence parsing to decompose the text into a structure from which relationships can be readily extracted (Friedman et al., 2001; Jelier et al., 2005). Among these approaches, rule learning is a useful data mining technique for discovering from high throughput biomedical data while manual annotation becomes more difficult. Therefore, there is an increasing need to automate the process. Moreover, rules have several advantages, including that they are easy for humans to interpret, represent knowledge modularly and can be applied using tractable inference procedures. Consequently, referring to the tools used in the work of Kim et al. (2007) the authors apply the rule-learning method for protein annotation, we pay attention to gene–disease relationship discovery in this study.

2. Architecture overview

Fig. 1 illustrates the overall architecture of our methods to the rule extraction in gene–disease relationship discovery. First, we preprocess each abstract in the corpus. We make use of some specific resources in the biomedical domain (e.g., Genia Tagger (http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/) and OMIM) in this phase. Next, we get a set of positive sentences and a set of negative sentences for learning. We refer to the morbid data from OMIM because it records some correct relationships between genes and diseases. Therefore, the morbid data can be regarded as a gold standard when evaluating the testing results. Then a memory-based shallow parser (http://www.clips.ua.ac.be/pages/MBSP#server) is adapted to produce the tagged sentences with the parsing information. After that, the ALEPH system (Srinivasan, 2000) is used to learn the relationships between genes and diseases. In the following, we select some rules according to the proposed methods and make rule paraphrase. Finally, a set of rules is obtained.

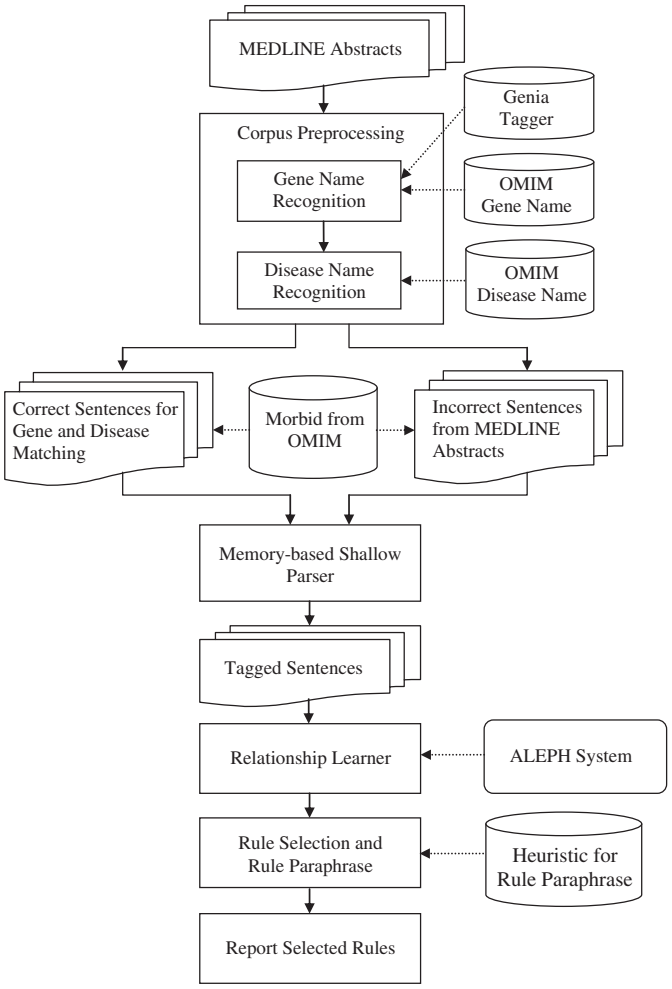

Fig. 1. System architecture.

3. Experimental data

As mentioned before, MEDLINE is a massive biomedical corpus and biomedical experts often retrieve special topics or extract associated documents from MEDLINE. It covers topics in biology, botany, biochemistry, biotechnology, medicine, nursing, dentistry, veterinary medicine, the health care system, and related fields. The partial documents used in TREC, 2004 Genome Track are considered as our experimental data (http://ir.ohsu.edu/genomics/2004protocol.html) which consist of 918,202 MEDLINE abstracts, involving a lot of genes from different species.

4. Methods

4.1. Corpus preprocessing

Our preprocessing procedure for each abstract consists of (1) gene name recognition and (2) disease name recognition. The details are explained as follows.

4.1.1. Gene name recognition

We use geniatagger.3.0.1 to identify all appearances of gene names. Since we are interested in human genes, we gather gene names from OMIM database and tag them in the experimental data to more completely annotate gene names. Genia Tagger developers report that the tagger has a precision of 67.45% and a recall of