Neurocomputing 206 (2016) 73–80

Contents lists available at ScienceDirect

Neurocomputing

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

A text feature-based approach for literature mining of lncRNA–protein interactions

Ao Li [a,b], Qiguang Zang [a], Dongdong Sun [a], Minghui Wang [a,b,*]

[a] School of Information Science and Technology, University of Science and Technology of China, 443 Huangshan Road, Hefei 230027, China
[b] Centers for Biomedical Engineering, University of Science and Technology of China, 443 Huangshan Road, Hefei 230027, China

ARTICLE INFO

Article history:
Received 12 August 2015
Received in revised form
30 October 2015
Accepted 20 November 2015
Available online 31 May 2016

Keywords:
LncRNA–protein interaction
Text mining
Text features
Machine learning

ABSTRACT

Long non-coding RNAs (lncRNAs) play important roles in regulating transcriptional and post-transcriptional levels. Currently, Knowledge of lncRNA and protein interactions (LPIs) is crucial for biomedical researches that are related to lncRNA. Many freshly discovered LPIs are stored in biomedical literature. With over one million new biomedical journal articles published every year, just keeping up with the novel finding requires automatically extracting information by text mining. To address this issue, we apply a text feature-based text mining approach to efficiently extract LPIs from biomedical literatures. Our approach consists of four steps. By employ natural language processing (NLP) technologies, this approach extracts text features from sentences that can precisely reflect the real LPIs. Our approach involves four steps including data collection, text pre-processing, structured representation, features extraction and training model and classification. The *F*-score performance of our approach achieves 79.5%, and the results indicate that the proposed approach can efficiently extract LPIs from biomedical literature.

1. Introduction

Long non-coding RNAs (lncRNAs) are defined as transcripts larger than 200 nt in length with limited protein-coding potential. Biologic studies have shown that lncRNAs play widespread roles in biologic processes by diverse mechanisms [1]. Interactions between lncRNA and protein now have been demonstrated to play important roles in transcriptional and post-transcriptional [2]. LncRNAs interact with protein such as chromatin-modification proteins and transcription factors. LncRNAs-related databases lncRNAdb [3] and the lncRNA and disease database [4] contain manually-collected lncRNA-protein interactions (LPIs) that reported on biomedical literatures. However, the speedy growth of the biomedical literature and the inefficient process has made the task of manual annotation increasingly difficult to achieve. We search lncRNA-related keywords such as lncRNA and lncRNA-protein in PubMed database, and count the number of relevant articles. Fig. 1 shows the number of the relevant article is increasing year by year. In particular, the number of relevant articles published in 2015 has reached about 70% of the total of 2014 at the moment of organizing this article. As a consequence, identification of LPIs from biomedical literature is an essential task in bioinformatics studies.

Currently, there are several text mining methods to extract protein-protein interactions (PPIs) and drug-drug interactions (DDIs) from biomedical text [5,6]. Methods have been proposed ranging from co-occurrence to machine learning (ML) combine with natural language processing (NLP) techniques [7]. The simplest method is co-occurrence [8], which achieves high recall but low precision. Conversely, approaches based on rule and pattern can improve precision but achieve greatly lower recall [9]. Furthermore, these rules or patterns are generated from training dataset, but rules- or patterns-based method often has low performance when we evaluate it based on test dataset. In addition, biomedical literature has diverse writing styles that make it difficult to find PPIs and DDIs by pattern matching. Therefore, the machine learning algorithm is recently incorporated to enhance the performance of PPIs and DDIs identification method. Many ML-based algorithms employed NLP techniques such as parser, which is always used in text mining research, for example, extraction protein phosphorylation from biomedical literature [10]. In this case, the PPIs and DDIs extraction work can be converted into a binary classification problem. A sentence that contains protein or drug pair can be represented by a set of text features (e.g. Lexical features, Phrase features, Verb features, etc.) [11], which are extracted from structured representation sentences.

[*] Corresponding author at: School of Information Science and Technology, University of Science and Technology of China, 443 Huangshan Road, Hefei 230027, China.
E-mail address: mhwang@ustc.edu.cn (M. Wang).

http://dx.doi.org/10.1016/j.neucom.2015.11.110
0925-2312/© 2016 Elsevier B.V. All rights reserved.

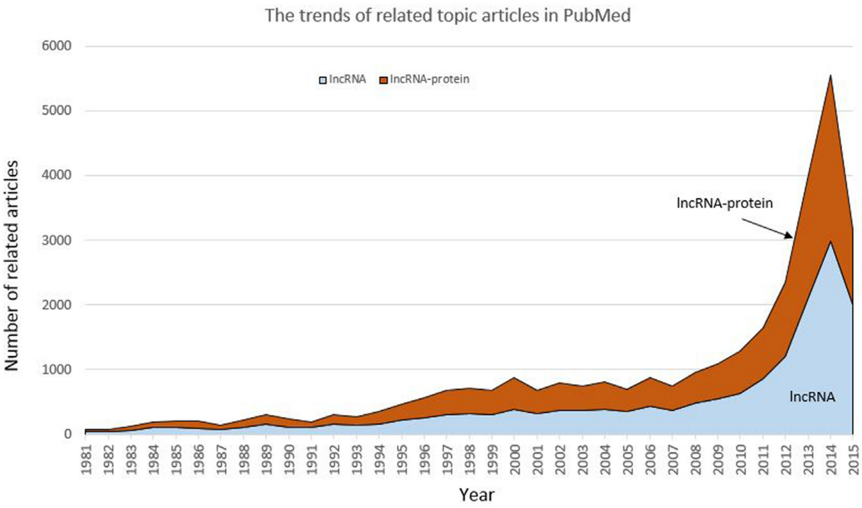

Fig. 1. The number of related topic articles in PubMed database. Blue and orange represent the number of lncRNA-related and lncRNA protein-related topic articles, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

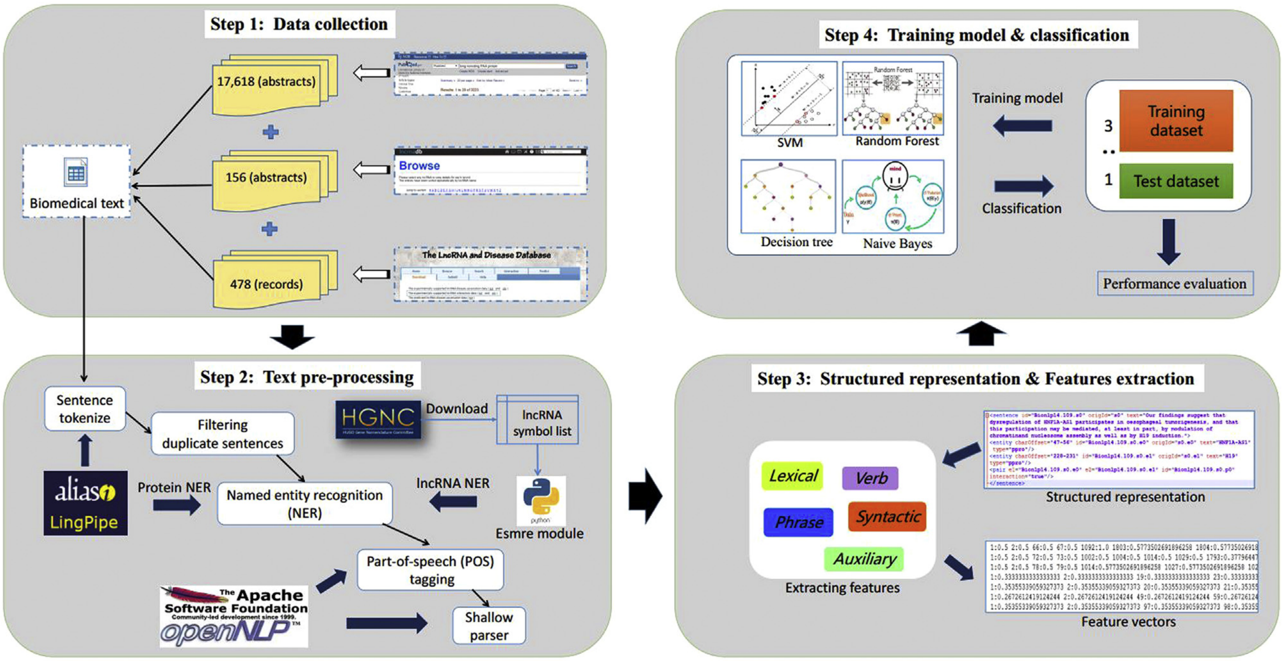

Fig. 2. Workflow of the proposed method. (Step 1) Download raw data from three databases and Unified text format, (Step 2) text pre-processing, (Step 3), structured representation and feature extraction, and (Step 4) Training model and classification.

Then these features are combined and represented as a feature vector. A machine learning algorithm is utilized to classify positive and negative instances based on feature vectors. Positive instances indicate that the sentence contains PPI or DDI. On the contrary, negative instances represent sentence do not include lncRNA or protein names, or sentence containing lncRNA and protein but do not indicate the interaction between lncRNA and protein. Several studies about automatic relation (e.g. PPIs, DDIs, gene-disease [5,6,12]) extraction is implemented successfully by text mining approach. LncRNA, protein, gene, disease and other terminology in the sentence of biomedical Literature has a similar composition and location. Therefore, automatic relation extraction methods can be effectively applied to LPIs extraction from the biomedical literature.

In this work, we introduce, to the best of our knowledge, the first text mining approach to automatically extract LPIs from literature. First, the method generates feature vectors by using a feature set including Lexical features, Phrase features, Verb features, Syntactic features, Auxiliary features, that have been proposed in previous studies [11]. Second, we utilize train dataset to train a binary supervised machine learning classifier. Afterwards, we analyze and evaluate our classifier model base on test dataset. This paper is structured as follows: Section 2 introduces the method of LPIs extraction in this work. Section 3 describes relevant experiments and analyses the results. Section 4 is the conclusion of this work, along with ideas for future work.

2. Materials and methods

The workflow of our method is illustrated in Fig. 2. Our method consists of three steps. Firstly, we collect raw data from three