Artificial Intelligence in Medicine 64 (2015) 185–193

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

Protein–protein interaction identification using a hybrid model

Yun Niu*, Yuwei Wang

College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, 29 Yudao Street, Qinhuaiqu, Nanjing, Jiangsu 210016, China

ARTICLE INFO

Article history:
Received 15 May 2014
Received in revised form 13 May 2015
Accepted 15 May 2015

Keywords:
Relational similarity model
Word similarity model
Biomedical text mining
Protein–protein interaction

ABSTRACT

Background: Most existing systems that identify protein–protein interaction (PPI) in literature make decisions solely on evidence within a single sentence and ignore the rich context of PPI descriptions in large corpora. Moreover, they often suffer from the heavy burden of manual annotation.
Methods: To address these problems, a new relational-similarity (RS)-based approach exploiting context in large-scale text is proposed. A basic RS model is first established to make initial predictions. Then word similarity matrices that are sensitive to the PPI identification task are constructed using a corpus-based approach. Finally, a hybrid model is developed to integrate the word similarity model with the basic RS model.
Results: The experimental results show that the basic RS model achieves *F*-scores much higher than a baseline of random guessing on interactions (from 50.6% to 75.0%) and non-interactions (from 49.4% to 74.2%). The hybrid model further improves *F*-score by about 2% on interactions and 3% on non-interactions.
Conclusion: The experimental evaluations conducted with PPIs in well-known databases showed the effectiveness of our approach that explores context information in PPI identification. This investigation confirmed that within the framework of relational similarity, the word similarity model relieves the data sparseness problem in similarity calculation.

1. Introduction

Information on protein–protein interactions (PPIs) is crucial for understanding the functional role of individual proteins as well as the entire biological process. Although numerous PPIs have been manually curated into database such as BioGRID [1], BIND [2],DIP [3], HPRD [4], IntAct [5] and MINT [6] by experts, information about many PPIs is still only available through the PubMed database. However, the amount of biomedical literature in PubMed grows rapidly and it is not practical to get complete coverage by manual curation. Therefore, mining PPIs from literature has become increasingly important and has attracted a lot of research interests. The well-known BioCreAtIvE (Critical Assessment of Information Extraction Systems in Biology) challenge includes a PPI detection task in two evaluations [7,8]. The primary goal of the task is to determine whether two target proteins interact.

Approaches for mining PPIs from biomedical text range from co-occurrence analysis to more sophisticated natural language processing systems. Co-occurrence analysis is the most straightforward approach and generally results in high recall but low precision [9,10]. Some other approaches construct patterns specifying how an interaction is described in literature and use them as rules to find PPIs [11–16]. Rule or pattern-based approaches can increase precision but significantly lower recall. In addition, these rule sets are derived from training data and are therefore not always applicable to other data they are not developed for [17,18]. In recent years, more and more approaches explore natural language processing technologies with a favor on machine learning (ML) methods. Some approaches focus on identifying features that are helpful in PPI identification, including lexical features, syntactic features, and semantic features [19–24]. Some approaches investigate various strategies of measuring the distance of two data points and explore it in kernel functions [25–31]. These ML approaches do not require manual construction of rules or patterns and often achieve better accuracy. However, they are experiencing some difficulties.

Given two target proteins, these ML approaches determine whether they interact based on evidence within a rather small text span, typically a sentence in which the proteins co-occur. Similar to other information extraction tasks, for PPI identification, the task is defined as determining whether there is an interaction relation between any two proteins mention in a sentence, as in the following example.

* Corresponding author. Tel.: +86 25 84896490.
E-mail address: yniu@nuaa.edu.cn (Y. Niu).

http://dx.doi.org/10.1016/j.artmed.2015.05.003
0933-3657/

*The screen identified interactions involving **c-Cbl** and two 14-3-3 isoforms, **cytokeratin 18**, human unconventional myosin IC, and a recently identified SH3 domain containing protein, **SH3 P17**.*

In this sentence, three proteins are mentioned (marked as bold). The task is to determine whether there is an interaction between any two of them, i.e., which ones of the three pairs (c-Cbl, cytokeratin 18), (c-Cbl, SH3 P17), (cytokeratin 18, SH3 P17) are interactions. The decision is made solely on evidence within this sentence.

These single-sentence-based approaches have some disadvantages. Firstly, complex syntactic structures of sentences often make the predictions very difficult. PPIs are complex biological processes and it is often the case that multiple proteins playing various roles are mentioned in the same sentence. Actually, in the AImed dataset [11] of PubMed abstracts annotated by human experts with protein interactions, over 40% of the sentences have more than three protein mentions. In order to depict these roles, complex syntactic structures are often used in a sentence. As a result, the connections of two protein mentions are often implicit, which makes it difficult to determine their relationship. As in the above sentence, there is a long distance (in terms of words) between c-Cbl and SH3 P17 and it would be difficult to derive a direct relation between them even through a deep syntactic parsing of the sentence. Secondly, context of interactions is ignored in these approaches. Actually, information in nearby sentences often provide the context of the interactions, thus could be very helpful in identifying the target interactions. However, this context is ignored in single-sentence-based approaches. In addition, an interaction may be reported and described by different pieces of research work hence appears in various papers. All these descriptions provide valuable evidence in recognizing the target PPI. Yet this information is not fully explored in the single-sentence-based approaches. Thirdly, these ML approaches suffer from small training datasets. In a single-sentence-based approach, in order to build the training data, every protein pair appearing in a sentence has to be manually annotated as positive (interactions) or negative (non-interactions). This is very intensive labeling work. As a result, these machine learning algorithms are usually trained on small datasets. This will inevitably affect the accuracy and portability of the models.

To address these issues, we propose a novel approach exploring corpus-based strategy to identify PPIs. Although there have been attempts to explore corpus-wide properties, they mostly explore frequency of interesting patterns [32,33]. Different from them, in the present work, relations between proteins are analyzed within the framework of relational similarity (RS) in natural language processing. In addition, a word similarity model that is derived from a large corpus is introduced to further improve the accuracy of the similarity calculation. Our method takes known PPIs in existing PPI databases (e.g., HPRD) as training data and no extra annotation is required. The experimental results show that this approach achieves high accuracy and well-balanced precision and recall.

The rest of this paper is organized as follows: Section 2 introduces the relational similarity framework. The process of PPI identification using the basic RS model and the results are discussed in Section 3. In Section 4, we introduce the word similarity model to further improve the accuracy of PPI identification and analyze the results of the hybrid model in detail. Section 5 concludes all our work.

2. The relational similarity framework

Research on relational similarity (RS) in the field of natural language processing provides a unified framework for accurately recognizing relations in text. Medin, Goldstone, and Gentner [34] describe relations as follows: relations are predicates taking two or more arguments (e.g., X collides Y, X is larger than Y), which are used to express abstract connection between objects. Most work on RS analysis tries to identify relations implied by word pairs, through comparing the similarity of the target relation with some known relations [35–38]. Usually, distributional properties of relations are first extracted from large-scale text. These properties characterize the connections between the two involved words. Then, some similarity measures are applied to calculate the similarity between the target relation and the known relations. The most similar one would be used to label the relation between the two target words.

Our decision to perform PPI recognition within the RS framework is based on two evaluations. First of all, interactions between proteins are typical semantic relations that match Medin's definition. More important, as discussed in the previous section, context information in a large corpus is crucial in determining whether two proteins interact. Within the RS framework, relations are indeed characterized by properties presented in large-scale text. This matches well with our intension to incorporate context in PPI recognition. Therefore, in the presented work, we analyze PPIs from the viewpoint of relational similarity. In the proposed method, the prediction is made upon the rich context information in a large corpus.

The RS framework contains three modules: collecting relation descriptions, relation representation, and similarity calculation. The first module is to get the collection of text that is likely to describe the relation between the two arguments from a large corpus. These descriptions can be phrases, sentences or paragraphs, etc. For example, Turney [35] selected 128 groups of phrases (e.g., X of Y, Y for X, X to Y) that contain the arguments (X, Y), while Nakov [36] used the set of sentences containing the two arguments. In the module of relation representation, vector space models are often used. Dimensions of the vectors correspond to properties characterizing the target relation. In the third module, appropriate similarity measures need to be designed and applied to calculate the distance between the target relation and the known relations. Finally, the target relation is labeled with the most similar known relation.

3. The basic relational similarity model in PPI recognition

3.1. System architecture

In the presented PPI recognition system, if two proteins interact, they form a *positive* pair. Otherwise, it is a *negative* pair. In order to determine whether two proteins interact, we calculate the similarity between the target pair and the known positive pairs, and the similarity between the target pair and the known negative pairs, respectively. The target pair gets a positive label if it is more similar to the positive pairs and a negative label otherwise.

Fig. 1 shows the architecture of the PPI identification system. The basic RS model is presented in the solid-line frame. As in the RS framework, our system of PPI recognition contains three modules, marked by the dashed-line frames in Fig. 1. They are described in the following subsections. The word similarity model is in the wavy-line frame and will be discussed in Section 4.

3.2. Collecting relation descriptions

The whole PubMed is used as the corpus from which descriptions of protein pairs are extracted. For a protein pair (p1, p2), we extract from PubMed all the sentences in which p1 and p2 co-occur, as they are likely to describe the relationship between p1 and p2. This set of sentences is regarded as the *signature* of (p1, p2). The signature is obtained by two steps.