Available online at www.sciencedirect.com

ScienceDirect

Expert Systems with Applications 33 (2007) 1036–1041

Expert Systems with Applications

www.elsevier.com/locate/eswa

Discovering gene–gene relations from sequential sentence patterns in biomedical literature

Jung-Hsien Chiang [a,*], Hsiao-Sheng Liu [b], Shih-Yi Chao [a], Cheng-Yu Chen [a]

[a] Department of Computer Science and Information Engineering, National Cheng Kung University, 1 Da-Shuei Road, Tainan 701, Taiwan
[b] Department of Microbiology and Immunology, College of Medicine, National Cheng Kung University, Tainan, Taiwan

Abstract

In this paper, we have developed a gene–gene relation browser (DiGG) that integrates sequential pattern-mining and information-extraction model to extract from biomedical literature knowledge on gene–gene interactions. DiGG combines efficient mining technique to enable the discovery of frequent gene–gene sequences even for very long sentences. Our approach aims to detect associated gene relations that are often discussed in documents. Integration of the related relations will lead to an individual gene relation network. Graphic presentation will be used to demonstrate the relationships between gene products. A salient feature of this approach is that it incrementally outputs new frequent gene relations in an online visualization fashion.

Keywords: Text mining; Bioinformatics; Sequential pattern mining; Information extraction; Gene networks

1. Introduction

Now that the Human Genome Project has completely accumulated sequences of human genes, the most challenging research has begun. The next step in genome analysis requires not only defining the function of each gene, but also determining the role of its interactions with other genes. In particular, the study of gene–gene interactions forms the basis for understanding the phenomena of activation, inhibition, down-regulation, up-regulation, and so on. Gene–gene interaction resources have been collected in databases such as MIPS, EcoCyc, and KEGG, but most are still not cataloged: information about them exists only in scientific literature, which is written in natural language that computers cannot easily understand. Efficient processing of large amounts of text to obtain this biological knowledge therefore requires sophisticated information extraction methods.

A number of methods have been proposed to generate patterns of information extraction in biomedical documents (Marcotte, Xenarios, & Eisenberg, 2001; Ono, 2001), for example, hand-coded pattern sets and statistical measures of keywords. Hand-coded pattern sets are based on significant interaction verbs and gene names, for example, [*Protein A interacts with Protein B*]. Such patterns yield fairly high precision but low recall, because there are many ways to express biological knowledge in natural language. Manually generated patterns are unreliable because there are many possible linkages between gene terms. Other methods are based on statistical measures of co-occurrence of keywords or gene names. This approach achieves high recall but low precision because it assumes that any pair of genes encountered in the same sentence interact, which is not always true. Many false-positives are thus retrieved because significant interaction keywords and gene names may occur in the same sentences when the genes mentioned are not syntactically related.

* Corresponding author. Tel.: +886 6 275 7575x62534; fax: +886 6 274 7076.
E-mail address: jchiang@mail.ncku.edu.tw (J.-H. Chiang).

0957-4174/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2006.08.017

Here are several examples from various biomedical documents of sentences that describe gene–gene relations:

1. "*In vitro experiments demonstrated that* ***MMP-9*** *was directly inhibited by* ***NAC*** *but was not influenced by* ***TPA.***" (*Anticancer Research, 21*(1A), 213–219)
2. "*At the same time,* ***PMA*** *induced hyperphosphorylation of* ***MARCKS*** *and* ***talin.***" (*International Journal of Cancer, 75*(5), 774–779)
3. "*Complex formation with the* ***MDM2*** *oncogene product is one mechanism inactivating the* ***p53 protein.***" (*European Urology, 32*(4), 487–493)
4. *Balance between activated-* ***STAT*** *and* ***MAP kinase*** *regulates the growth of human bladder cell lines after treatment with epidermal growth factor.* (*International Journal of Oncology, 15*(4), 661–667)

It can be seen from above examples that the syntactic relationships between words can be *positive* or *negative*. A positive syntactic relationship (e.g. *induce, inhibit, inactivate, regulate*) characterizes the G–G relations in sentence, while a negative one (e.g. *not, but, and, nor*) signals no or even reversed relations. A syntactic relationship must be positive in order to determine what sort of G–G relation exists. Moreover, active (or passive) description also expresses an ordered sequence. These sequences represent true biological relations in gene products. In this study, we use a sequential-pattern-mining algorithm to identify interaction patterns between genes. In specific, we propose a sequential mining-based hybrid model to mine meaningful information-extraction rules that delineate the kinds of morphological features that can appear before and after the gene names in sentences describing gene–gene interactions in documents. This interaction identification traditionally demands heavy resources and often includes extensive cross-referencing.

2. Methods

2.1. System architecture

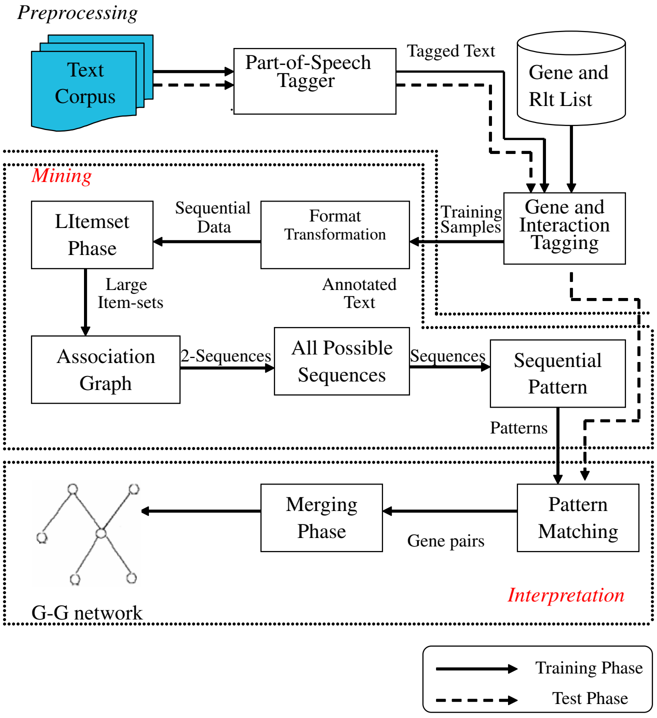

Fig. 1. Schematic diagram of the DiGG system.

Scientific literature carries much information. To make that information easily and efficiently accessible to researchers, the literature must be computer-readable and causality-interpretable. One way this can be done is by first dividing each document into its constituent sentences and then using a shallow parser to identify the part-of-speech (POS) of each word in individual sentence. The parsing results can then be used as training samples for the subsequent sequential pattern-mining algorithm. The mining stage is for finding candidate frequent sequences within those sentences. The mining algorithm is especially efficient when the sequential gene/relation patterns in the database are complicated. Fig. 1 shows a schematic flow diagram of the proposed method, which consists of three components in the DiGG system: *the preprocessing stage, the mining stage, and the interpretation stage*. The approach can be summarized as follows:

1. tagging the parts of speech and gene names/relations,
2. extracting gene–gene interaction rules,
3. mining all positive syntactic patterns from the training samples,
4. associate candidate sequences,
5. display the evidence of possible gene–gene relationships graphically.

In this section, we discuss the detailed procedures for the proposed framework and briefly describe the developed system, DiGG (Fig. 2).

2.2. Mining information extraction rules

In this study, we are interested in the biologically sequential relations between genes, not in the words used to describe those relations. We therefore need to divide sentences into several blocks based on stopwords, gene names, and relational terms. A valid sentence will be transformed into time-sequential data from left to right. Fig. 3 illustrates how a preprocessed training sample is divided into several blocks. Look at the following training-sample sentence:

"IL-6/*Gene* was/*vbd* found/*vbn* to/*to* decrease/*Rlt* mdr1/*Gene*".