Discovering Gene-Gene Relations from Fuzzy Sequential Sentence Patterns in Biomedical Literature

Jung-Hsien Chiang*, Zong-Xian Yin and Cheng-Yu Chen
Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan 70101
* E-mail: jchiang@mail.ncku.edu.tw

Abstract

We have developed a gene-gene (G-G) relation browser that combines fuzzy sequential pattern mining and information-extraction model to extract from biomedical literature knowledge on gene-gene interactions. Our approach aims to detect associated G-G relations that are often discussed in documents. Integration of the related relations will lead to an individual G-G network. Graphic presentation will be used to demonstrate the relationships between gene products.

I. 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, 2001; Ono, 2002; Chiang, 2004], 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.

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

- "In vitro experiments demonstrated that **MMP-9** was directly inhibited by **NAC** but was not influenced by **TPA**." (Anticancer Res 21(1A):213-9)
- "At the same time, **PMA** induced hyperphosphorylation of **MARCKS** and **talin.**" (Int J Cancer, 75(5):774-9)
- "Complex formation with the **MDM2** oncogene product is one mechanism inactivating the **p53 protein**." (Eur Urol, 32(4):487-93)
- Balance between activated **STAT** and **MAP kinase** regulates the growth of human bladder cell lines after treatment with epidermal growth factor.(Int J Oncol, 15(4):661-7)

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 fuzzy sequential-pattern-mining algorithm to identify interaction patterns between genes. In specific, we propose a sequential mining-based hybrid fuzzy 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.

II. Methods

A. System architecture

Scientific literature carries much information. To make that information easily and efficiently accessible to researchers, the literature must be computer-readable. 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 of each word in each sentence. The parsing results can then be used as training samples for the subsequent sequential pattern-mining algorithm. Figure 1 shows a schematic flow diagram of the proposed method, which consists of three components in the proposed system: *the preprocessing stage, the mining stage, and the interpretation stage*.

In the preprocessing stage, parts of speech and gene name/relations are tagged. In the mining stage, gene-gene interaction rules are extracted and all positive syntactic patterns from the training samples are found by the sequential pattern-mining technique. In the interpretation stage, the evidence of possible gene-gene relationships is displayed graphically (Fig. 2).

B. 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. Look at the following training-sample sentence:

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

According to our gene lexicon, "IL-6" and "mdrl" are marked as gene names; based on a relation dictionary, "decrease" falls into the "relation (*Rlt*)" category. Those words that are not included in the categories of gene names and relation words will retain their original part-of-speech designation, for example, *"vbd"*, *"vbn"*, and *"to"* above. We then place these training blocks into modified sequential pattern-mining algorithms [Yen, 1996] to obtain G-G relation-extraction rules. The proposed sequential pattern mining algorithms not only discover large itemsets (a group of items that appear together), but also identify large sequences (an ordered list of sets of items). The ordered list includes the patterns of gene-gene interactions, i.e. shows exactly which gene acts on which gene (or genes).

III. Results and discussions

We utilized our system to search arsenic-induced bladder-cancer-related genes and their relations [Simeonova, 2000; Sanchez-Carbayo, 2003]. A total of 9,870 corresponding abstracts were retrieved from the PubMed database. We then automatically identified human gene names from those abstracts to filter out non-relevant documents. We applied our system to these documents and found 48 sequential G-G relations through valid sentences. Thirty-nine G-G relation descriptions were confirmed as correct by medical experts. Examples of correct relation pairs included "*PMA* <up-regulate> *VEGF*", "*LPA* <induce> *ACTIN*", "*p53* <activate> *WAF1*", etc. The precision of our system was 81.3% and the recall rate was 73%. According to Staab [2002], the precision rate of manual information-retrieval techniques applied to biomedical documents is above 90% but with a recall rate of less than 20%. The reasons for this already described above. The recall rate for the proposed system is reasonable because of fuzzy mining algorithm's "fault tolerance" in extracting interaction rules. A limitation of our system is that it cannot handle the sort of very long and broadly descriptive terms that have been popularized by the biomedical community. This requires more research. Two false-positive examples in which the proposed system is unable to identify the gene-gene relations are shown below.

- After incubation with 4-ABP, F-actin decreased and G-actin increased in both cytoplasm and nuclei of PC cells and cytoplasmic F-actin fibers were lost, but only cytoplasmic actin was altered in the BC cells.

- Utilizing ASO directed against the raf-1 gene, a central component of this proposed pathway, we were able to reverse the RR phenotype of human tumor cell lines having elevated HER-2 expression or a mutant form of Ha-ras, two genes upstream of raf-1 in signal transduction.

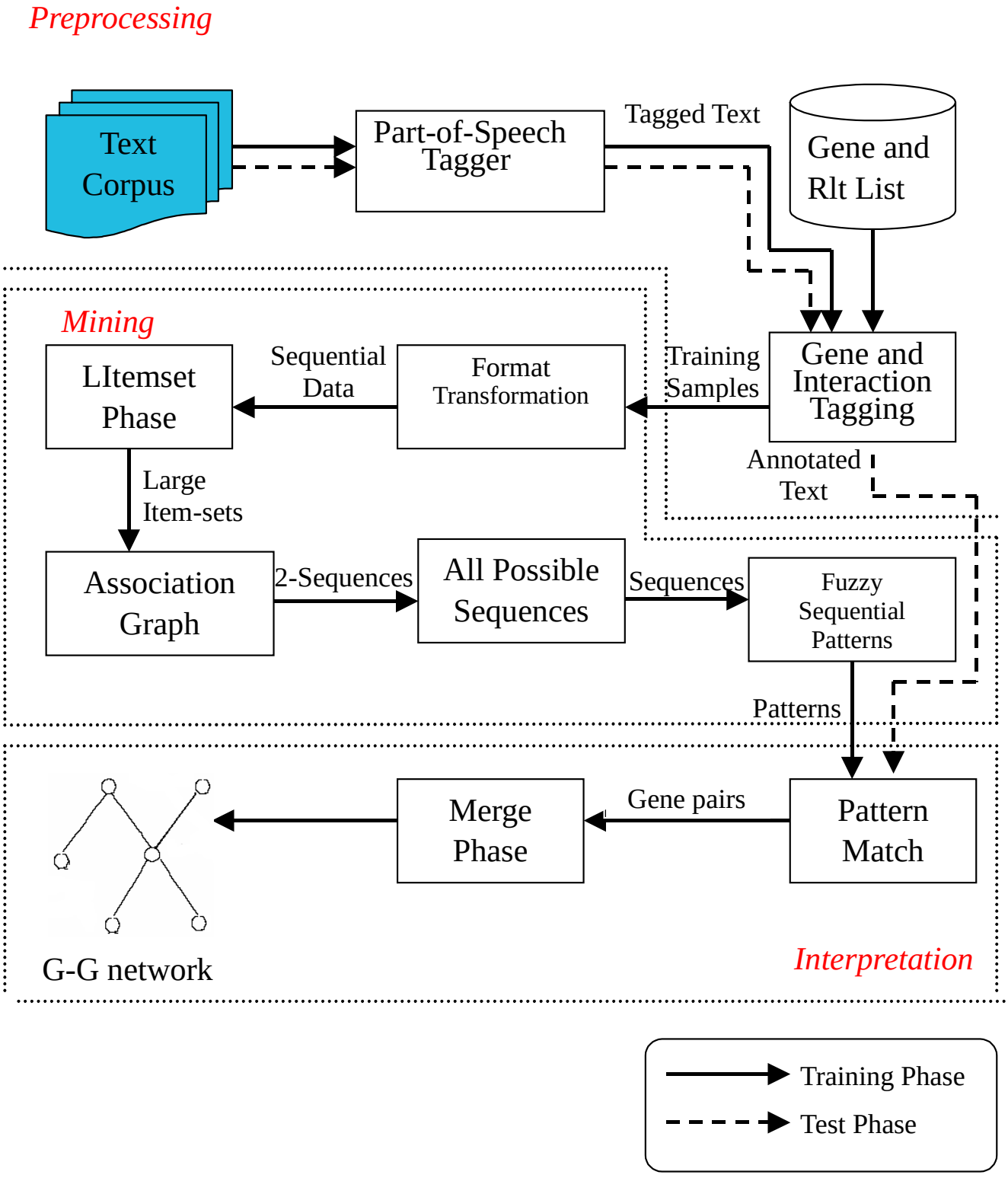

Figure 1 Schematic diagram of the proposed system.

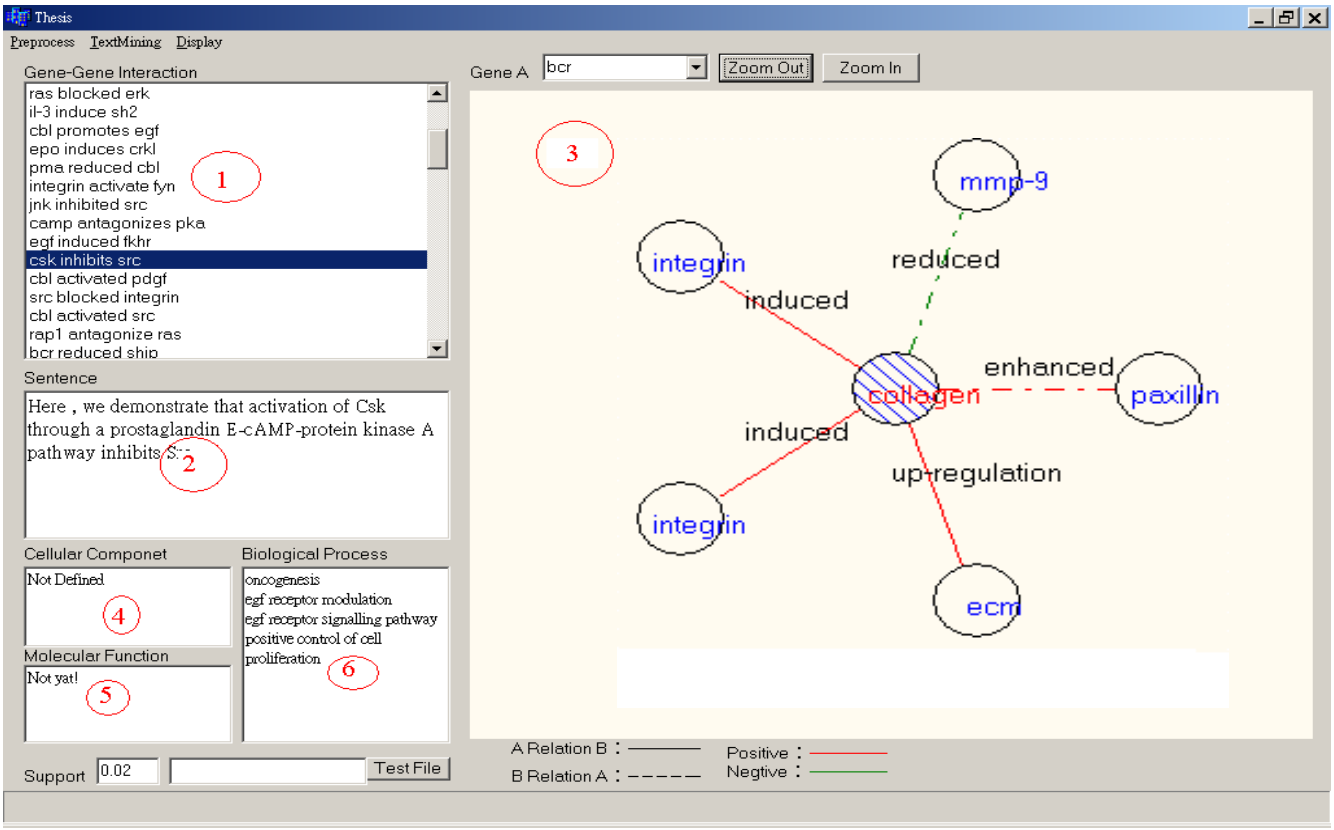

Figure 2 Graphic interface of the proposed system. View #1 lists relations extracted from biomedical documents. When users click on one of those relations, it will be able to read its original description in View #2. The G-G network in View #3 displays valid relations between gene products. When users move the cursor to certain node of a gene in View #3, the system will show related information concerning its cellular component, molecular function and biological process from the LocusLink in View 4, 5 and 6, respectively.

IV. References

[1] Chiang J. -H., H. -C. Yu, and H. -J. Hsu; GIS: a biomedical text-mining system for gene information discovery. Bioinformatics, 20 (1), 120-121, 2004.

[2] Marcotte E. M., L. Xenarios, and D. Eisenberg; Mining literature for protein-protein interactions. Bioinformatics, 17, 359-363, 2001.

[3] Ono T.; Automated extraction of information on protein-protein interactions from the biological literature, Bioinformatics, 17, 155-161, 2001.

[4] Sanchez-Carbayo M. and C. Cordon-Cardo; Applications of array technology: identification of molecular targets in bladder cancer. British Journal of Cancer, 89, 2172-2177, 2003.

[5] Simeonova P. et al.; Arsenic mediates cell proliferation and gene expression in the bladder epithelium: association with activating protein-1 transactivation. Cancer Research, 60, 3445-3453, 2000.

[6] Staab S.; Mining information for functional genomics. IEEE Intelligent System, 17, 70-73, 2002.

[7] Yen S. J. and A. L. P. Chen; An efficient approach to discovering knowledge from large databases**.** 4th International Conference on Parallel and Distributed Information Systems (PDIS '96), 8-18, December 18, 1996.