Interdiscip Sci Comput Life Sci (2009) 1: 179–186
DOI: 10.1007/s12539-009-0028-7

Text Mining Biomedical Literature for Constructing Gene Regulatory Networks

Yong-Ling SONG[1], Su-Shing CHEN[1,2*]
[1](Department of Computer and Information Science and Engineering, University of Florida Gainesville, FL 32611 USA)
[2](CAS-MPG Partner Institute of Computational Biology, Shanghai Institutes of Biological Sciences, Chinese Academy of Sciences, Shanghai, 200031 China)

Received 18 October 2008 / Revised 6 December 2008 / Accepted 7 December 2008

Abstract: In this paper, we present the framework of a Gene Regulatory Networks System: GRNS. The goals of GRNS include automatically mining biomedical literature to extract gene regulatory information (strain number, genotype, gene regulatory relation, and phenotype), automatically constructing gene regulatory networks based on extracted information and integrating biomedical knowledge into the regulatory networks.

Key words: gene regulatory network, text mining.

1 Introduction

Through decades of active research, tremendous amounts of experimental data are available on the gene function and their regulation in different genomes (Goodman *et al.*, 2004; Woods, 2004). However, from the experimental data, embedded in tens of thousands of published literature, it is difficult for the individual researcher to extract a comprehensive view of the gene function and their regulation in different genomes. Furthermore, rapid progress of the research on different genomic sequences within recent years brings in exponential growth of related literature (Afantenos *et al.*, 2005; Cohen *et al.*, 2004; Cohen *et al.*, 2005; Hirschman *et al.*, 2003; Liu *et al.*, 2003; Nenadic *et al.*, 2003; Shatkay *et al.*, 2003; Yandell *et al.*, 2003). To help individual scientists to keep up with all the new information, a complete system that not only compiles the experimental evidences but also logically integrates the knowledge related to gene function and regulation is desired.

In recent years, the extraction of knowledge from biological literature has received considerable attention. For example, Blaschke *et al.* (1999) used the statistical "bag of words" approach to the extraction protein-protein interaction. Yakushiji *et al.* (2001) designed an information extraction system using a general-purpose full parser. Friedman *et al.* (2001) presented a GENIES system which extracts structured information about cellular pathways from biomedical literatures. Marcotte *et al.* (2001) showed a Bayesian approach of mining literature. McDonald *et al.* (2004) developed an Arizona Relation Parser for extracting gene pathway relations. Chun *et al.* (2005) introduced a system to extract disease-gene relations from Medline by using a dictionary matching with machine learning-based named entity recognition approach. Hu *et al.* (2005) developed a rule-based system RLIMS-P to do the database annotation of protein phosphorylation. Yuan *et al.* (2006) developed a web-based version of RLIMS-P. Saric *et al.* (2005) presented a rule based approach for extracting information from biomedical text.

There are two most used methods to extract biological knowledge: either a statistical method based on co-occurrences of proteins or genes, or a rule-based extraction method. Statistical methods are good at locating potential protein-protein interactions. But, they usually cannot provide a clear classification of interaction information. Rule-based relation extraction methods can achieve good precision and recall if the manually developed pattern is good. For example, Hu *et al.* (2005) developed very good pattern templates to extract protein phosphorylation information. But it is difficult to build a set of complete pattern templates even for a biological expert. In this paper, we provide a framework of a rule-based method with the help of potential informative sentences discovering. Therefore, new templates and rules can be incrementally supplemented.

2 System Overview

2.1 System Architecture of GRNS

Figure 1 shows the system architecture of GRNS

*Corresponding author.
E-mail: suchen@picb.ac.cn

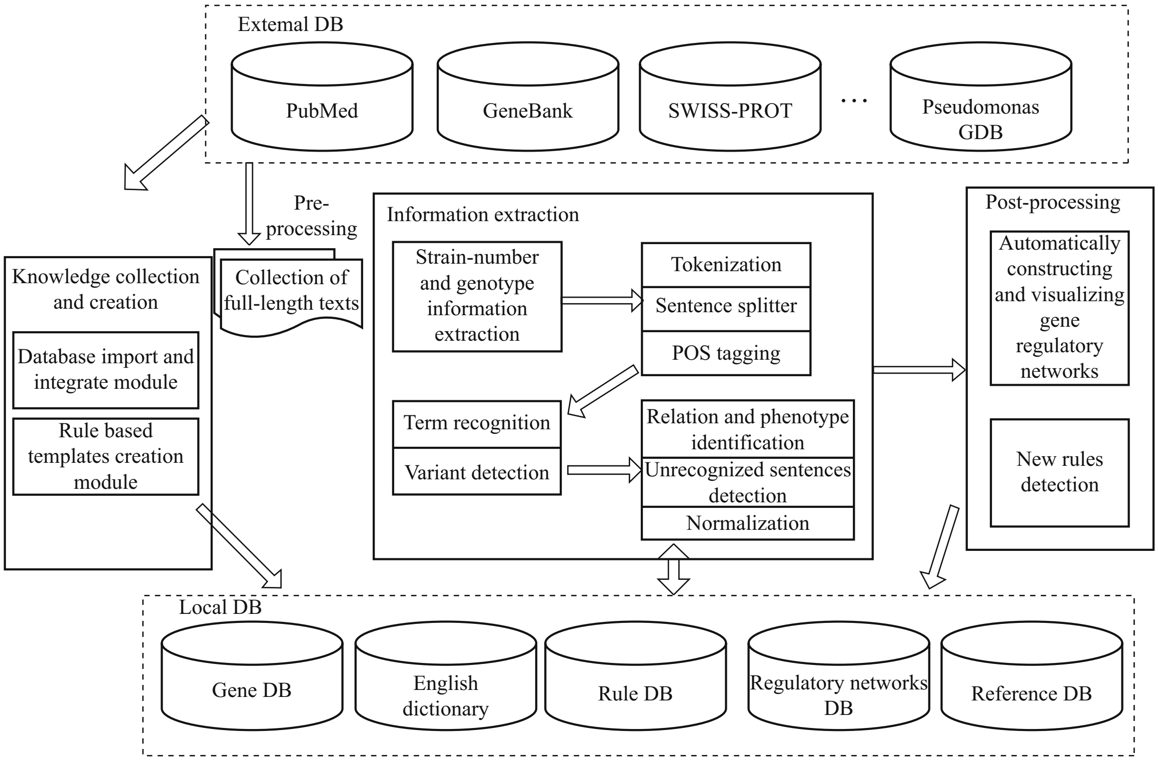

Extemal DB
PubMed
GeneBank
SWISS-PROT
...
Pseudomonas GDB
Pre-processing
Knowledge collection and creation
Database import and integrate module
Rule based templates creation module
Collection of full-length texts
Information extraction
Strain-number and genotype information extraction
Tokenization
Sentence splitter
POS tagging
Term recognition
Variant detection
Relation and phenotype identification
Unrecognized sentences detection
Normalization
Post-processing
Automatically constructing and visualizing gene regulatory networks
New rules detection
Local DB
Gene DB
English dictionary
Rule DB
Regulatory networks DB
Reference DB

Fig. 1 GRNS System Architecture

system. GRNS consists of four modules, Knowledge Collection and the Creation Module, the Pre-Processing Module, the Information Extraction Module (IE for short), and the Post-Processing Module. The external database resource includes some general database resources, such as GeneBank and SWISS-PROT (Boeckmann *et al.*, 2003). It also includes some specific organism resources, for example, Pseudomonas aeruginosa Genome Database (Stover *et al.*, 2000) for Pseudomonas aeruginosa Genome. The tasks of the Knowledge Collection and the Creation Module include: first, integrating external database data into the local database; second, creating the templates and rules knowledge for the Information Extraction Module. The Pre-Processing Module downloads the full-length biomedical texts from PubMed Database and sends the corpus of texts to the IE Module. First the IE Module does strain-number and genotype information extraction. Then IE Module does the tokenization, sentences splitter and Part of Speech tagging. Then the IE Module recognizes the gene, the protein entities, and discriminating words from the corpus of texts. Then it extracts the relation, the phenotype, and other kinds of entities based on a rule-based approach. Finally, the extracted information is normalized based on the normalized rule. After the IE Module processes the collected text, the Post-Processing Module first automatically constructs the regulatory networks based on the extracted information and specific existing knowledge, such as gene functional classes' knowledge and subsystem knowledge. Then the Post-Processing Module saves potential informative sentences into the database. Experts can browse these potential sentences and may create new kinds of rules for later use.

2.2 Data Modeling of GRNS

The main objective of GRNS is to automatically extract the gene regulatory information from a collection of unstructured biomedical text. Here, the biomedical text is any research paper. Usually these papers are downloaded from the PubMed database. But what is the definition of the gene regulatory information, and what kind of data are we interested in the information extracting? Basically, GRNS extracts five kinds of data after processing the biomedical text: the gene regulatory relation information, the strain number, the genotype, the phenotype, and unrecognized sentences. Why we need these five kinds of data? Clearly, to construct a gene regulatory network, we need the gene regulatory relation information. The regulation can be at the transcriptional level (activation or repression), the posttranscriptional (mRNA stability) level, the translational level or the post-translational (protein-protein interaction/modification) level. We need the strain number, the genotype, and the phenotype information for the data reliability reason. All this information is part of the evidence of gene regulation. They help the researcher to validate the gene regulatory relation information. For unrecognized sentences, it helps us to discover new rules in the information extraction module.

2.3 Visualization Modeling of Gene Regulatory Networks

GRNS automatically provides the visualization of regulatory networks. To do this, we provide visualization modeling of regulatory networks. Visualized regulatory networks include two kinds of information, entities and relations. In GRNS, there are seven kinds of entities: genes, proteins, operons, products, merged-genes, subsystem, and step. An operon is a group of key nucleotides sequences that are controlled and usually function as a unit. GRNS has simple operons and complex operons. In a simple operon, there is no relationship between operon's genes. In a complex operon, there are relationships between the operon's genes. Merged-genes are not natural genes and they are used to improve the layout. Step is used to describe the biomedical process. A subsystem usually means a group of related functional roles which are jointly involved in a specific aspect of the cellular machinery. In GRNS, there are more than ten kinds of relations between entities, such as DNA binding, RNA binding, protein binding, the two-component regulatory system, the signal molecule production, signal sensing, the product, the signal/molecule binding, activate, required for and repress. We provide notations and symbols for visualization modeling of the regulatory networks to generate interactive graphical regulatory networks for subsystems or the whole genome. Figure 2 illustrates these notations and symbols, including three kinds of information: entities, relations, and color information for different function classes.

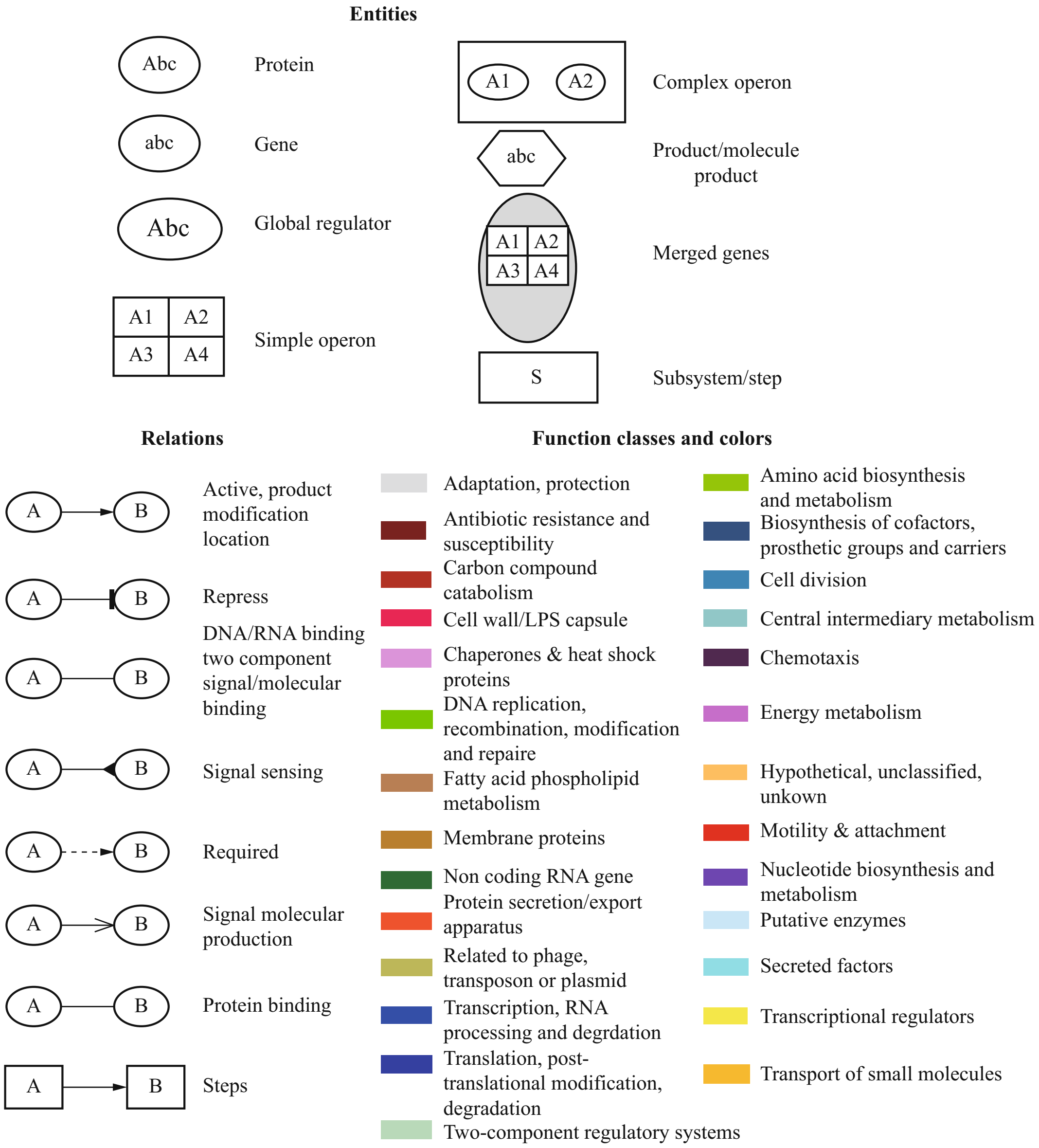

Entities
Abc
Protein
abc
Gene
Abc
Global regulator
A1 A2 A3 A4
Simple operon
A1 A2
Complex operon
abc
Product/molecule product
A1 A2 A3 A4
Merged genes
S
Subsystem/step
Relations
A B
Active, product modification location
A B
Repress
A B
DNA/RNA binding two component signal/molecular binding
A B
Signal sensing
A B
Required
A B
Signal molecular production
A B
Protein binding
A B
Steps
Function classes and colors
Adaptation, protection
Antibiotic resistance and susceptibility
Carbon compound catabolism
Cell wall/LPS capsule
Chaperones & heat shock proteins
DNA replication, recombination, modification and repaire
Fatty acid phospholipid metabolism
Membrane proteins
Non coding RNA gene
Protein secretion/export apparatus
Related to phage, transposon or plasmid
Transcription, RNA processing and degrdation
Translation, post-translational modification, degradation
Two-component regulatory systems
Amino acid biosynthesis and metabolism
Biosynthesis of cofactors, prosthetic groups and carriers
Cell division
Central intermediary metabolism
Chemotaxis
Energy metabolism
Hypothetical, unclassified, unkown
Motility & attachment
Nucleotide biosynthesis and metabolism
Putative enzymes
Secreted factors
Transcriptional regulators
Transport of small molecules

Fig. 2 Visualization Annotation and Symbols for GRNS.

3 Methods

3.1 Information Extraction (IE)

3.1.1 Strain Table Analysis

The strain number and genotype are important gene regulatory information. Usually, there is a table in the biomedical paper showing all the strains used in the study and their genotypes. We analyze the strain table and extract the genotype and strain number from the table. One difficulty in analyzing the strain table is that we do not have a structured table data directly. We need to extract the table information from the unstructured text. Usually, we do not know where the table starts, when one table column ends, when one table row ends, and when the table ends.

Fortunately, most biomedical papers provide the strain table in a standard format. We find out that there are some rules to help us to recognize the strain tables. For example, to find where the strains table start, we found most of the strain table names include "strains" or "plasmids", such as, "Bacterial strains and plasmids used in this study", "Strains, plasmids, and primers used in this study", "Bacterial strains and plasmids". If we get one line text like "Table...strains...(plasmids)...", usually it means a strain table. Another example, when we need to recognize how many columns in this table and what the meaning of each column is, we found in most cases, these columns are also in a standard format. Most tables include three columns, the first one for the strain number, the second one for the genotype, and the last one for the reference information. To recognize these columns, we can follow these patterns. The column name for the strain number usually includes strain or plasmid, such as, "Strain or plasmid", "Strain, plasmid, or oligonucleotide". The column name for genotype usually includes genotype or description or characteristics, such as, "Description or sequence", "Characteristics" or "Description". The column name for reference usually includes the reference or the source, such as, "Reference", "Source or Reference". Based on this information, we can recognize the column names. To recognize the strain number and genotype information, we also found there are some rules to help us. For example, we find out that most strain numbers are one word. If the strain number is more than one word, usually the first word of the multi-words strain number is also a strain number itself. One example of multi-words strain number is "PAO1 ncr". Here PAO1 itself is a strain number. So, we can analyze the strains table using all these rules.

3.1.2 Gene, Protein and Discriminating words Recognition

After strains table analysis, the IE Module first performs the tokenization, sentences splitter and Part of Speech tagging. To tag the words with POS labels, it uses the Brill part-of-speech tagger (Brill, 1995). After POS tagging, IE performs term recognition and variant detection to recognize discriminating words and Gene/Protein names. Discriminating words recognition is relatively simple: given the pre-defined words set, with the detection of synonym and different verbal form, we can detect the synonym with the help of a synonym dictionary and label the discriminating words with any verbal form. Recognizing the Gene/Protein names is challenging. We need to take care of the problems of the expanded form of abbreviation, homology and aliases. To deal with these problems, we use a gene-dictionary of aliases and abbreviation. We construct the gene-dictionary by combining multi-database recourses, such as the SWISS_PROT and the *Pseudomonas aeruginosa* Genome Database.

3.1.3 Relation and Phenotype Identification

We use the cascaded finite state automata to recognize the gene regulatory relation and phenotype information. The cascaded finite state automata are implemented by a CASS parser (Abney, 1996). CASS parser is a robust and speedy partial parser. Our rules for gene regulatory relation and phenotype information recognition are written in the CASS grammar. Some previous systems also use the CASS parser to recognize the regulatory gene/protein relation information. The most famous one is the STRING-IE system in EMBL project (Saric *et al.*, 2005). We follow some basic grammar in STRING-IE with some supplemental grammar rules. For example, in Tables 1, and 2, there are some examples of our CASS grammar in finding the gene regulatory relations. These examples cannot be recognized by the STRING-IE CASS grammar. The first line of these tables is the grammar we provide for the CASS parser. The next line is the meaning of this grammar. Then we provide a real sentence from a biomedical paper. In the end, we provide the extracted gene regulatory relation information.

Table 1 CASS grammar example 1.

Grammar	ex_reg− >nxpg (cma?wdt?) (rb)* (cma?neg?) (vx?) (rb)* (VERB) ownexpr
Explanation	One possible gene relation is gene A regulate its own expression.

3.1.4 Unrecognized Sentences Detection and Ranking

It is difficult to build up complete rules to recognize all entities and relations. IE stores the potential sentences to detect possible template candidates. Experts can create new templates based on template candidates' information. We choose the template candidates in this way: if a sentence includes gene/protein names and fails to match the existing pattern, we assign this sentence as a template candidate. However, it is difficult to discover useful information if there are too many template candidates. Therefore, we need to rank the template candidates. Ling *et al.* (2006) provided a sentence-ranking schema which is very good for evaluating sentences. We use this schema with some changes. Our sentence-ranking schema includes two kinds of ranking:

(1) Document Relevance Score-Sd

A document relevance score uses the similar vector space model and cosine similarity to compare a sentence vector and document vector. Details of Sd can be found in (Ling *et al.*, 2006).

(2) Location Document Relevance Score-Sl

In (Ling *et al.*, 2006), the schema assigned the location score as follows: if the sentence is the last sentence of an abstract, then 1, otherwise 0. We find there are other location-related hints in evaluating the sentence. First, title information is usually important and informative, so we assign title's location score as 1. Second, we find that whenever one gene or protein is mentioned in the text, the description of its function or relationship with other gene/subsystem may be somewhere close. So, if a sentence includes a gene/protein name, and this name is mentioned nearly (at most three sentences away), we assign Sl as 1.

(3) Sentence-Ranking Score

The final score of a sentence S is a weighted sum of the three scores: S=0.7 Sd+0.3 Sl. After IE Module, GRNS stores all candidate sentences and their sentences-ranking scores, and sorts all sentences based on sentence-ranking scores.

Table 2 CASS grammar example 2.

<table>
<tr><td>Grammar</td><td>locate_ge ne (vx?) (ADV)* (vx?) (rb)* (VERB) (in |of| by) nxpg;</td></tr>
<tr><td>Explantion</td><td>Gene A located in ups tream/downstream gene B relate gene C.</td></tr>
<tr><td>Example sentence</td><td>fim S located immediately ups tream of algR is also reau ired for twithing motilitv</td></tr>
</table>

3.2 Automatic Construction and Visualization of Regulatory Networks

After the IE Module, the GRNS constructs the regulatory networks based on the entity and relation information extracted from biomedical literature. GRNS provides an automatically interactive visualization method to visualize and integrate the biomedical evidence to the visualized regulatory networks. Interactive visualization of regulatory networks provides an interactive way to browse the regulatory networks (Wu *et al.*, 2005).

4 Results and Evaluation

First, we show the text mining results for the paper—"Biosynthesis of Pyochelin and Dihydroaeruginoic Acid Requires the Iron-Regulated pchDCBA Operon in *Pseudomonas aeruginosa*" in the GRNS (PMID: 8982005) (Serino *et al.*, 1997). Then we provide a regulatory network constructed by extracting data from 200 randomly selected papers about the *Pseudomonas aeruginosa* Genome and filter out to display the data on the Type III secretion subsystem. *Pseudomonas aeruginosa* is an environmental bacterium, which causes serious human infections, especially in those with reduced immunity, patients with Cystic Fibrosis or severe burns (Shiwani *et al.*, 1997). In the end, we provide the evaluation result for the GRNS.

4.1 A text mining results for the paper with PMID 8982005

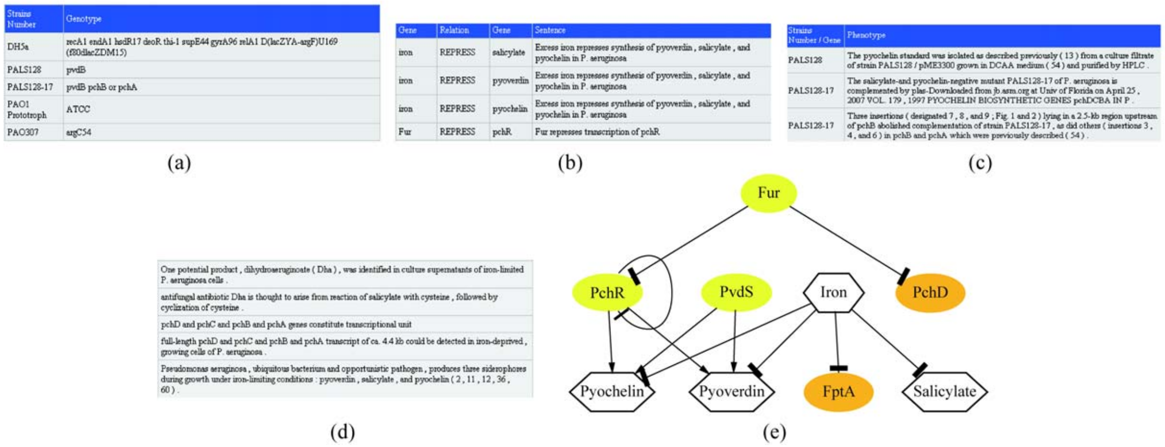
The text mining results for the paper with PMID 8982005 is shown in Fig. 3. The results include the strain number and genotype information, the gene regulatory relation information, the phenotype information, the unrecognized sentences, and the visualized gene regulatory network based on the extracted gene relation information.

4.2 The Type III secretion Subsystem

The regulatory network of Type III secretion subsystem (TTSS for short) is shown in Fig. 4. The TTSS is an important virulence factor of *P. aeruginosa*: it inhibits host defense systems by inducing apoptosis in macrophages, polymorphonuclear phagocytes, and cells. The TTSS contains a syringe like apparatus, which can directly inject the effector proteins from the bacterium cytoplasm into the host cell cytosol, causing cell death. The *P. aeruginosa* TTSS machinery is encoded by 31 genes arranged in four operons on the chromosome. Four effector proteins, ExoS, ExoT, ExoY and ExoU have been found in *P. aeruginosa*. According to the current working model, the needle forms a pole in the host cell membrane, and the effector proteins are delivered through the hollow needle. Based on published research, we divide the TTSS translocation process into six steps, as presented in the black boxes. Following the boxes, we use the "Type III Secretion System" to represent the overall function of this subsystem. We find that the regulatory network can clearly describe the relationship in the Type III secretion subsystem.

4.3 The Evaluation Results

We use precision and recall to evaluate the results of the GRNS. Precision and recall are the most common parameters when evaluating the IE system. Precision is to evaluate whether the system can only extract correct information, recall is to evaluate whether the system can recognize all useful information (Cohen, 2005). The



Strains Number	Genotype
DH5a	recA1 endA1 hsdR17 deoR thi-1 supE44 gyrA96 relA1 D(lacZYA-argF)U169 (f80dlacZDM15)
PALS128	pvdB
PALS128-17	pvdB pchB or pchA
PAO1 Prototroph	ATCC
PAO307	argC54

(a)

Gene	Relation	Gene	Sentence
iron	REPRESS	salicylate	Excess iron represses synthesis of pyoverdin , salicylate , and pyochelin in P. aeruginosa
iron	REPRESS	pyoverdin	Excess iron represses synthesis of pyoverdin , salicylate , and pyochelin in P. aeruginosa
iron	REPRESS	pyochelin	Excess iron represses synthesis of pyoverdin , salicylate , and pyochelin in P. aeruginosa
Fur	REPRESS	pchR	Fur represses transcription of pchR

(b)

Strains Number / Gene	Phenotype
PALS128	The pyochelin standard was isolated as described previously (13) from a culture filtrate of strain PALS128 / pME3300 grown in DCAA medium (54) and purified by HPLC .
PALS128-17	The salicylate-and pyochelin-negative mutant PALS128-17 of P. aeruginosa is complemented by plas-Downloaded from jb.asm.org at Univ of Florida on April 25 , 2007 VOL. 179 , 1997 PYOCHELIN BIOSYNTHETIC GENES pchDCBA IN P .
PALS128-17	Three insertions (designated 7 , 8 , and 9 ; Fig. 1 and 2) lying in a 2.5-kb region upstream of pchB abolished complementation of strain PALS128-17 , as did others (insertions 3 , 4 , and 6) in pchB and pchA which were previously described (54) .

(c)

One potential product , dihydroaeruginoate (Dha) , was identified in culture supernatants of iron-limited P. aeruginosa cells .
antifungal antibiotic Dha is thought to arise from reaction of salicylate with cysteine , followed by cyclization of cysteine .
pchD and pchC and pchB and pchA genes constitute transcriptional unit
full-length pchD and pchC and pchB and pchA transcript of ca. 4.4 kb could be detected in iron-deprived , growing cells of P. aeruginosa .
Pseudomonas aeruginosa , ubiquitous bacterium and opportunistic pathogen , produces three siderophores during growth under iron-limiting conditions : pyoverdin , salicylate , and pyochelin (2 , 11 , 12 , 36 , 60) .

(d)

(e)

Fig. 3 A text mining result for the paper with PMID 8982005 (a) strain number and genotype information (b) the gene regulatory relation information (c) the phenotype information (d) the unrecognized sentences (e) visualized gene regulatory network.

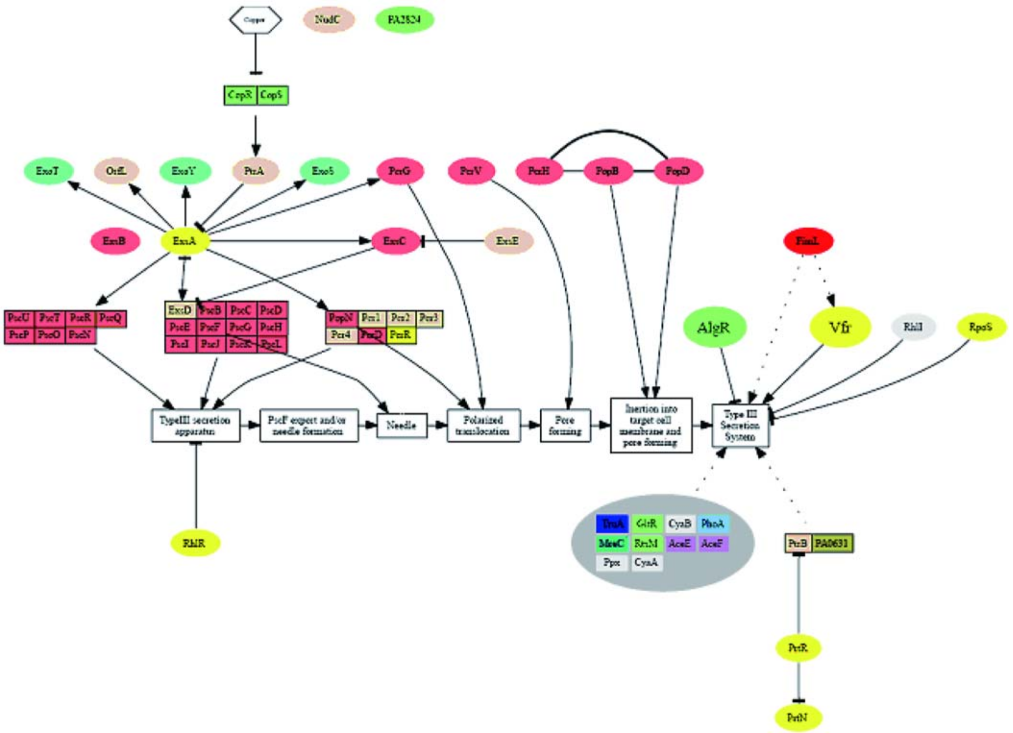


Fig. 4 The type III secretion subsystem

definition of precision and recall is shown as follows:

$$\text{Precision} = \text{numberofcorrectlyextractedentities}/\text{numberoftotalextractedentities}$$

$$\text{Recall} = \text{numberofcorrectlyextractedentities}/\text{numberofallcorrectentities}$$

To evaluate the precision and recall of the extracted information, it is necessary to manually analyze all information in the corpus and compare them with the extracted information. We provide the evaluation results in Table 3. From the evaluation results in Table 3, we can see that the GRNS has high precision and good recall in extracting relation information and other gene regulatory information.

Table 3 Evaluation result for GRNS.

Name	Precision	Recall
Strain number	0.93	0.92
Genotype	0.90	0.89
Gene Regulatory Relation	0.91	0.79
Phenotype	0.87	0.74

In the above table, the recall rate was lower 0.74, because we had to build a more elaborate dictionary and grammar for the specific PA genome in order to improve the results. In our future work, we will improve these results.

5 Conclusion

In this paper, we presented the framework of a Gene Regulatory Networks System: GRNS. It can automatically mine biomedical literature and construct gene regulatory networks. GRNS utilizes a rule-based method to extract useful information from biomedical literature. Then GRNS automatically constructs and visualizes the regulatory networks based on the information extracted and some existed domain-specific knowledge. To supplement the manual templates, GRNS detects the potentially informative sentences and save them. All saved sentences are sorted by a heuristic sentence ranking score.

References

[1] Abney, S. 1996. Statistical Methods and Linguistics. In: The Balancing Act: Combining Symbolic and Statistical Approaches to Language. The MIT Press, 1–26.

[2] Afantenos, S., Karkaletsis, V., Stamatopoulos, P. 2005. Summarization from medical documents: a survey. Artificial Intelligence in Medicine 33, 157–177.

[3] Blaschke, C., Andrade, M.A., Ouzounis, C., Valencia, A. 1999. Automatic extraction of biological information from scientific text: protein-protein interactions. In: Proceedings International Conference on Intelligent Systems in Molecular Biology 7, 60–67.

[4] Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I. 2003. The SWISS-PROT protein knowledgebase and its supplement TREMBL. Nucleic Acids Research 31, 365–370.

[5] Brill, E. 1995. Transformation-based error-driven learning and natural language pro-pessing: A case study in part-of-speech tagging. Computational Linguistics 21, 543–566.

[6] Chun, H., Tsuruoka, Y., Kim, J., Shiba, R., Nagata, N., Hishiki, T., Tsujie, J. 2005. Etraction of Gene-Disease relations from medline using domain dictionaries and maching learning. In: Proceedings of Pacific Symppsium Biocomputing 11, 4–15.

[7] Cohen, A.M., Hersh, W.R. 2005. A survey of current work in biomedical text mining. Briefings in Bioinformatics 6, 57–71.

[8] Cohen, K., Hunter, L. 2004. Natural language processing and systems biology. In: Dubitzky and Pereira (eds), Artificial intelligence methods and tools for systems biology, Springer Verlag, 147–173.

[9] Friedman, C., Kra, P., Yu, H., Krauthammer, M., Rzhetsky, A. 2001. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. Bioinformatics 17 (Suppl 1), S74–82.

[10] Goodman, L., Lory, S. 2004. Analysis of regulatory networks in Pseudomonas aeruginosa by genomewide transcriptional profiling. Current Opinions on Microbiology 7, 39–44.

[11] Hirschman, L., Park, J.C., Tsujii, J., Wong, L., Wu, C.H. 2003. Accomplishments and challenges in literature data mining for biology. Bioinformatics 18, 1553–1561.

[12] Hu, Z.Z., Narayanaswamy, M., Ravikumar, K.E., Vijay-Shanker, K., Wu, C.H. 2005. Literature mining and database annotation of protein phosphorylation using a rule-based system. Bioinformatics 21, 2759–2765.

[13] Ling, X., Jing, J., He, X., Mei, Q., Zhai, C., Schatz, B. 2006. Automatically generating gene summaries from biomedical literature. In: Proceeding of Pacific Symposium Biocomputing, 11, 41–50.

[14] Liu, H., Friedman, C. 2003. Mining terminological knowledge in large biomedical corpora. In: Proceeding of Pacific Symposium Biocomputing 8, 415–426.

[15] Marcotte, E.M., Xenarios, I., Eisenberg, D. 2001. Mining literature for protein-protein interactions. Bioinformatics 17, 359–363.

[16] McDonald, D.M., Chen, H., Su, H., Marshall. B.B. 2004. Extracting gene pathway relations using a hybrid grammar: the Arizona Relation Parser. Bioinformatics 20, 3370–3378.

[17] Nenadic, G., Spasic, I., Ananiadou, S. 2003. Terminology-driven mining of biomedical literature. Bioinformatics 19, 938–943.

[18] Saric, J., Jensen, L.J., Ouzounova, R., Rojas, I., Bork, P. 2005. Extraction of regulatory gene / protein networks from Medline. Bioinformatics 22, 654–650.

[19] Serino, L., Reimmann, C., Visca, P., Beyeler, M., Chiesa, V.D., Haas, D. 1997. Biosynthesis of pyochelin and dihydroaeruginoic acid requires the iron-regulated pchDCBA operon in Pseudomonas aeruginosa. Journal of Bacteriology 179, 248–257.

[20] Shatkay, H., Feldman, R. 2003. Mining the biomedical literature in the genomic era: an overview. Journal of Computational Biolology 10, 821–855.

[21] Shiwani, A.K., Ritchings, B.W., Almina, E.C., Lory, S., Ramphal, R. 1997. A Transcriptional Activator, FleQ, Regulates Mucin Adhesion and Flagellar Gene Expression in Pseudomonas aeruginosa in a Cascade Manner. Journal of Bacteriology 179, 5574–5581.

[22] Stover, C.K., Pham, X.Q., Erwin, A.L., Mizoguchi, P., Warrener, M.J., Hickey, F.S., Brinkman, S.D., Hufnagle, W.O., Kowalik, D.J., Lagrou, M., Garber, R.L., Goltry, L., Tolentino, E., Westbrock-Wadman, S., Yuan, Y., Brody, L.L., Coulter, S.N., Folger, K.R., Kas, A., Larbig, K., Lim, R., Smith, K., Spencer, D., Wong, G.K., Wu, Z., Paulsen, I.T., Reizer, J., Saier,

M.H., Hancock, R.E., Lory, S., Olson, M.V. 2000. Complete genome sequence of Pseudomonas aeruginosa PA01, an opportunistic pathogen. Nature 406, 959–964.

[23] Woods, D.E. 2004. Comparative genomic analysis of Pseudomonas aeruginosa virulence. Trends Microbiology 12, 437–439.

[24] Wu, W., Song, Y., Jin S., Chen, S. 2005. An Interactive Map of Regulatory Networks of Pseudomonas aeruginosa Genome. In: Proceedings of First RECOMBS Satellite Workshop on Systems Biology, 1–10.

[25] Yakushiji, A., Tateisi, Y., Miyao, Y., Tsujii, J. 2001. Event extraction from biomedical papers using a full parser. In: Proceedings of Pacific Symposium Biocomputing, 408–419.

[26] Yandell, M.D., Majoros, W.H. 2003. Genomics and natural language processing. Nature Reviews Genetics 3, 601–610.

[27] Yuan, X., Hu, Z.Z., Wu, H.T., Torii, M., Narayanaswamy, M., Ravikumar, K.E., Vijay-Shanker, K., Wu, C.H. 2006. An online literature mining tool for protein phosphorylation. Bioinformatics 22, 1668–1669.