

# Genescene: Biomedical Text And Data Mining

Gondy Leroy<sup>1</sup>, Hsinchun Chen<sup>1</sup>, Jesse D. Martinez<sup>2</sup>, Shauna Eggers<sup>1</sup>, Ryan Falsey<sup>2</sup>, Kerri Kislin<sup>2</sup>, Zan Huang<sup>1</sup>, Jiexun Li<sup>1</sup>, Jie Xu<sup>1</sup>, Daniel McDonald<sup>1</sup>, Gavin Ng<sup>1</sup>

Artificial Intelligence Lab<sup>1</sup>

Department of Management Information Systems

The University of Arizona

Tucson, Arizona 85721, USA

1-520-621-2748

Arizona Cancer Center<sup>2</sup>

The University of Arizona

Tucson, Arizona 85721, USA

{gleroy@eller.arizona.edu; hchen@eller.arizona.edu; seggers@email.arizona.edu; jmartinez@azcc.arizona.edu; rfalsey@u.arizona.edu; kkislin@u.arizona.edu; zhuang@eller.arizona.edu; jiexun@eller.arizona.edu; jxu@eller.arizona.edu; dmm@eller.arizona.edu; tgavinng@ai.bpa.arizona.edu}

## ABSTRACT

To access the content of digital texts efficiently, it is necessary to provide more sophisticated access than keyword based searching. Genescene provides biomedical researchers with research findings and background relations automatically extracted from text and experimental data. These provide a more detailed overview of the information available. The extracted relations were evaluated by qualified researchers and are precise. A qualitative evaluation of the current online interface indicates that this form of searching is more useful and efficient than keyword based searching.

## Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval – *Information Filtering, Retrieval Models, Search Process, and Selection Process.*

## General Terms

Algorithms

## Keywords

Natural Language Processing, NLP, Deep Semantic Parsing, Biomedicine, Ontology, Rule-based, Corpus-based, Genescene.

## 1. INTRODUCTION

The Internet has increased the availability of and access to publications, leading in many cases to information overload. In biomedicine, this effect has been accelerated by an increase in publications and dataset due to the decoding of the human genome. Every 11 years, the number of researchers doubles [6] and Medline, the main resource of research literature, has been growing with over 10,000 abstracts per week since 2002 [5]. Simple keyword-based document retrieval is inadequate in this field. In addition, large amounts of gene expression data (typically involves expression measurements on thousands of genes) have been available from microarray experiments recently. Gene expression patterns embedded in such data have the potential to lead to discovery on unknown genetic relations. The

information overload problem for both literature search and data analysis calls for technical solutions to largely automate these processes.

Genescene, a toolkit developed for biomedicine, will provide more adequate access to text and data. It allows researchers to view the findings extracted from Medline abstracts, and compare them with findings from microarray experiments.

## 2. RELATED WORK

To access the content of documents, natural language processing (NLP) is required. In biomedicine, many NLP approaches start from a list with specific gene names or verbs and extract the surrounding text as relations [2; 7; 8]. The best achieve high precision but low recall, since few relations are extracted. Co-occurrence based approaches assume that phrases, e.g., genes, appearing in the same text are related [3]. This approach extracts more but less precise relations. Both ignore negation.

## 3. GENESCENE

Genescene combines relations between entities in text, extracted by a rule-based parser and a corpus-based co-occurrence analysis technique. It also extracts regulatory relations from microarray data and combines them with relations extracted from text.

### 3.1 Rule-based Findings

The rule-based parser, successfully tested with a small prototype [4], is based on closed-class words which provide a generic structure for the relations. Cascaded finite state automata (FSA), built around prepositions and basic sentence elements, identify the structures. For example, from the sentence “*MDM2 suppresses p73 function without promoting p73 degradation*,” the relation “*Mdm2 – suppress – p73 function*” is extracted. Each FSA incorporates negation, an important element ignored by others, and also captures relations based on conjunctions.

### 3.2 Corpus-based Background

The corpus-based background relations, represent the knowledge in the entire domain that forms the background for the rule-based relations, are formed by a co-occurrence-based algorithm tested earlier in an information retrieval context [1]. These are relations between noun phrases that hold true for the entire collection.

### 3.3 Ontologies

Three ontologies, the Gene Ontology (GO), the Human Genome (HUGO) Nomenclature, and the Unified Medical Language System (UMLS), are used to better integrate the relations. A term can receive semantic tags based on its being in the ontologies.

### 3.4 Regulatory Relations from Data

Regulatory relations are extracted from microarray data using data mining techniques such as Bayesian networks and association rule mining. Incorporating these regulatory relations into Genescene will help researchers compare experimental discovery with previous knowledge from literature. Unexpected gene associations can be identified to guide further literature search or new experimental design.

### 3.5 Online Access

Users access Genescene online and choose the type of relation they want to see (<http://ai.bpa.arizona.edu/go/genescene>). A list of relations is then shown for the search, ordered by the type and number of elements in the relation. The number of abstracts containing the relation is also shown. Clicking on a relation retrieves the list of associated abstracts and the abstracts themselves when required. The relations and search terms are highlighted in the abstracts.

## 4. CASE STUDY

Extracted relations and the online interface were evaluated with a quantitative and a qualitative study, respectively.

### 4.1 Genescene Relations

Two researchers evaluated the rule- and corpus-based relations from p53-related abstracts. The rule-based relations were 95% correct and the corpus-based relations were 60% correct. Limiting the corpus-based relations to those that had entities with ontology tags increased their correctness to 78%. Most terms and relations were considered relevant, especially when part of an ontology.

### 4.2 Genescene Content

Encouraged by the excellent evaluation results, three collections with biomedical abstracts were added to Genescene. Different research groups requested the collections. Table 1 provides an overview. The P53 collection contains all abstracts available in Medline (summer 2002) with the keyword "p53" in the title or abstract. The AP1 collection is based on the keywords: ap1, ap-1, jnk, erk, jun, fos, and p38. The yeast collection is based on "yeast." Hundreds of thousands of relations are available for each collection. In each, more than half of the terms received a tag from an ontology. The UMLS provided more than 50% of the terms with a tag. Slightly more than 1% of the terms received a

**Table 1. Overview of Genescene Content**

Topic:	P53	AP1	Yeast
Abstracts:	23,234	30,820	56,246
Rule-B. Relations:	270,008	387,666	560,165
Corpus-B. Relations:	5,023,103	6,526,454	7,736,647
Terms w. UMLS tag:	57%	54%	51%
Terms w. HUGO tag:	0.6%	0.9%	1.4%
Terms w. GO tag:	1.1%	1.1%	1.1%
Any ontology tag:	58%	55%	52%

GO-tag. HUGO provided the least tags, except for the yeast collection where 1.4% of the terms received a HUGO tag.

### 4.3 Genescene Interface

The ordering of the relations was very much liked. Researchers felt the most important relations, e.g., conclusions, were presented first. They also liked the highlighting of relations and keywords in the original abstracts. However, currently only one keyword or phrase can be used to search a collection. Multiple-keyword-search should be done. In addition, the speed of partial match queries for background relations should be improved.

## 5. FUTURE DIRECTIONS

In the future, we will integrate both types of relations from text and provide an interactive graphical map display. Later, visual text mining will become possible. Users will be able to search for specific time ranges in the publications or for specific organism. Furthermore, findings from text will be incorporated into microarray data mining to improve the algorithmic performance.

## 6. ACKNOWLEDGMENTS

Genescene is supported by the grant: NIH/NLM, 1 R33 LM07299-01, 2002-2005, "Genescene: a Toolkit for Gene Pathway Analysis." We thank the National Library of Medicine, the Gene Ontology Consortium, and the Hugo Nomenclature Committee for making the ontologies available to researchers.

## 7. REFERENCES

- [1] Chen, H. and Lynch, K. J. Automatic Construction of Networks of Concepts Characterizing Document Databases. *IEEE Transactions on Systems, Man and Cybernetics*, 22, 5 (1992), 885-902.
- [2] Friedman, C., Kra, P., Yu, H., Krauthammer, M. and Rzhetsky, A. GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, 17, Suppl. 1 (2001), S74-S82.
- [3] Jenssen, T.-K., Laegreid, A., Komorowski, J. and Hovig, E. A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, 28, (2001), 21-28.
- [4] Leroy, G. and Chen, H. *Filling Preposition-based Templates to Capture Information from Medical Abstracts*. Paper presented at the Pacific Symposium on Biocomputing, Kauai (2002), 350-361.
- [5] National-Library-of-Medicine, Fact Sheet - Medline. National Library of Medicine: <http://www.nlm.nih.gov/pubs/factsheets/medline.html>
- [6] Perutz, M. F. Will biomedicine outgrow support? *Nature*, 399, (1999), 299-301.
- [7] Pustejovsky, J., Castaño, J., Zhang, J., Kotecki, M. and Cochran, B. *Robust Relational Parsing Over Biomedical Literature: Extracting Inhibit Relations*. Paper presented at the Pacific Symposium on Biocomputing (2002), 362-373.
- [8] Thomas, J., Milward, D., Ouzounis, C., Pulman, S. and Carroll, M. *Automatic Extraction of Protein Interactions from Scientific Abstracts*. Paper presented at the Pacific Symposium on Biocomputing (2000), 538-549.