BioLINK 2004: Linking Biological Literature,
Ontologies, and Databases, pp. 69-76.
Association for Computational Linguistics

# Using Natural Language Processing, Locus Link, and the Gene Ontology to Compare OMIM to MEDLINE

**Bisharah Libbus** **Halil Kilicoglu** **Thomas C. Rindflesch** **James G. Mork** **Alan R. Aronson**

Lister Hill National Center for Biomedical Communications
National Library of Medicine
Bethesda, Maryland, 20894
{libbus|halil|tcr|mork|alan}@nlm.nih.gov

## Abstract

Researchers in the biomedical and molecular biology fields are faced with a wide variety of information sources. These are presented in the form of images, free text, and structured data files that include medical records, gene and protein sequence data, and whole genome microarray data, all gathered from a variety of experimental organisms and clinical subjects. The need to organize and relate this information, particularly concerning genes, has motivated the development of resources, such as the Unified Medical Language System, Gene Ontology, LocusLink, and the Online Inheritance In Man (OMIM) database. We describe a natural language processing application to extract information on genes from unstructured text and discuss ways to integrate this information with some of the available online resources.

## 1 Introduction

The current knowledge explosion in genetics and genomics poses a challenge to both researchers and medical practitioners. Traditionally, scientific reviews, which summarize and evaluate the literature, have been indispensable in addressing this challenge. OMIM (Online Mendelian Inheritance in Man) (OMIM 2000), for example, is a clinical and biomedical information resource on human genes and genetic disorders. It has close to 15,000 entries detailing clinical phenotypes and disorders as well as information on nearly 9,000 genes. The database can be searched by gene symbol, chromosomal location, or disorder.

More recently, automated techniques for information and knowledge extraction from the literature are being developed to complement scientific reviews. These methods address the need to condense and efficiently present large amounts of data to the user. The feasibility of applying natural language processing techniques to the biomedical literature (Friedman and Hripcsak 1999; de Bruijn and Martin 2002) and to the wealth of genomics data now available (Jenssen et al. 2001; Yandell and Majoros 2002) is increasingly being recognized. Efforts to develop systems that work toward this goal focus on the identification of such items as gene and protein names (Tanabe and Wilbur 2002) or groups of genes with similar function (Jenssen et al. 2001; Masys et al. 2001). Other groups are interested in identifying protein-protein (Blaschke et al. 1999; Temkin and Gilder 2003) or gene-gene interactions (Stephens et al. 2001; Tao et al. 2002), inhibit relations (Pustejovsky et al. 2002), protein structure (Gaizauskas et al. 2003), and pathways (Ng and Wong 1999; Friedman et al. 2001).

We discuss the modification of an existing natural language processing system, SemGen (Rindflesch et al. 2003), that has broad applicability to biomedical text and that takes advantage of online resources such as LocusLink and the Gene Ontology. We are pursuing research that identifies gene-gene interactions in text on genetic diseases. For example the system extracts (2) from (1).

1) Here, we report that TSLC1 directly associates with MPP3, one of the human homologues of a Drosophila tumor suppressor gene, Discs large (Dlg).

2) TSLC1|INTERACT_WITH|MPP3

Due to the complexity of the language involved, the extraction of such predications is currently not accurate enough to support practical application. However, we suggest its potential in the context of an application that combines traditional, human-curated resources such as OMIM and emerging information extraction applications.

## 2 Molecular Biology Resources

To support and supplement the information extracted by SemGen from biomedical text, we draw on two resources, LocusLink and the Gene Ontology. LocusLink (Wheeler et al. 2004) provides a single query interface to curated genomic sequences and genetic loci. It presents information on official nomenclature, aliases, sequence accessions, phenotypes, OMIM numbers, homology, map locations, and related Web sites, among others. Of particular interest is the Reference Sequence (RefSeq) collection, which provides a comprehensive, curated, integrated, non-redundant set of sequences, including genomic DNA, transcript (RNA), and protein products for major research organisms. Currently, SemGen uses LocusLink to obtain normalized gene names and Gene Ontology annotations.

The Gene Ontology (GO) (The Gene Ontology Consortium 2000, 2001, 2004) aims to provide a dynamic controlled vocabulary that can be applied to all organisms, even while knowledge of gene and protein function is incomplete or unfolding. The GO consists of three separate ontologies: molecular function, biological process, and cellular component. These three branches are used to characterize gene function and products and provide a comprehensive structure that permits the annotation of molecular attributes of genes in various organisms. We use GO annotations to examine whether there are identifiable patterns, or concordance, in the function of gene pairs identified by SemGen.

## 3 SemGen

SemGen identifies gene interaction predications based on semantic interpretation adapted from SemRep (Srinivasan and Rindflesch 2002; Rindflesch and Fiszman 2003), a general natural language processing system being developed for the biomedical domain. After the application of a statistically-based labeled categorizer (Humphrey 1999) that limits input text to the molecular biology domain, SemGen processing proceeds in three major phases: categorial analysis, identification of concepts, and identification of relations.

The initial phase relies on a parser that draws on the SPECIALIST Lexicon (McCray et al. 1994) and the Xerox Part-of-Speech Tagger (Cutting et al. 1992) to produce an underspecified categorial analysis.

In the phase for identifying concepts, disorders as well as genes and proteins are isolated by mapping simple noun phrases from the previous phase to concepts in the Unified Medical Language System® (UMLS)® Metathesaurus® (Humphreys et al. 1998), using MetaMap (Aronson 2001). ABGene, a program that identifies genes and proteins using several statistical and empirical methods (Tanabe and Wilbur 2002) is also consulted during this phase. In addition, a small list of signal words (such as *gene*, *codon*, and *exon*) helps identify genetic phenomena. For example, the genetic phenomena in (4) are identified from the sentence in (3). Concepts isolated in this phase serve as potential arguments in the next phase.

3) WIF1 was down-regulated in 64% of primary prostate cancers, while SFRP4 was up-regulated in 81% of the patients.

4) genphenom|WIF1
genphenom|SFRP4

During the final phase, in which relations are identified, the predicates of semantic propositions are based on indicator rules. These stipulate verbs, nominalizations, and prepositions that "indicate" semantic predicates. During this phase, argument identification is constrained by an underspecified dependency grammar, which also attempts to accommodate coordinated arguments as well as predicates.

SemGen originally had twenty rules indicating one of three etiology relations between genetic phenomena and diseases, namely CAUSE, PREDISPOSE, and ASSOCIATED_WITH. In this project, we extended SemGen to cover gene-gene interaction relations: INHIBIT, STIMULATE, AND INTERACT_WITH. About 20 indicator rules were taken from MedMiner (Tanabe et al. 1999). We supplemented this list by taking advantage of the verbs identified in syntactic predications by GeneScene (Leroy et al. 2003). SemGen has 46 gene-gene interaction indicator rules (mostly verbs), including 16 for INHIBIT (such as *block*, *deplete*, *down-regulate*); 12 for INTERACT_WITH (*bind*, *implicate, influence*, *mediate*); and 18 for STIMULATE (*amplify*, *activate, induce, up-regulate*).

An overview of the SemGen system is given in Figure 1, and an example is provided below. SemGen processing on input text (5) produces the underspecified syntactic structure (represented schematically) in (6). (7) illustrates genetic phenomena identified, and (8) shows the final semantic interpretation.

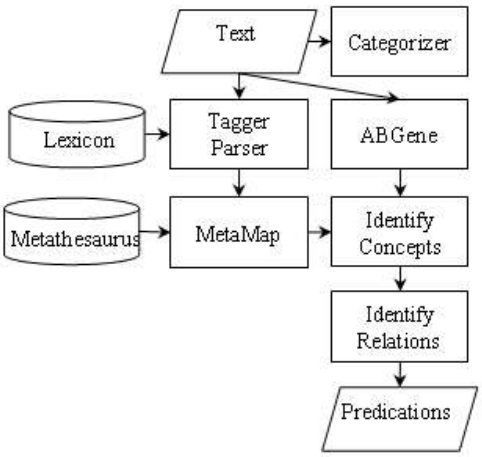


Figure 1. SemGen system

5) We show here that EGR1 binds to the AR in prostate carcinoma cells, and an EGR1-AR complex can be detected by chromatin immunoprecipitation at the enhancer of an endogenous AR target gene.

6) [We] [show] [here] [that] [EGR1] [binds] [to the AR] [in prostate carcinoma cells,] [and] [an EGR1-AR complex] [can] [be] [detected] [by chromatin immunoprecipitation] [at the enhancer] [of an endogenous AR target gene]

7) genphenom|egr1
genphenom|ar
genphenom|enhancer endogenous ar target gene

8) egr1|INTERACT_WITH|ar

During processing, SemGen normalizes gene symbols using the preferred symbol from LocusLink. The final interpretation with LocusLink gene symbol is shown in (9).

9) EGR1|INTERACT_WITH|AR

As we retrieve the LocusLink symbol for a gene, we also get the GO terms associated with that gene. We are interested in extending the application of our textual analysis and knowledge extraction methodology and relating it to other biomedical and genomic resources. Gene Ontology is one such important resource, and below we discuss the possibility that GO might shed additional light on the biological relationship between genes that are paired functionally based on textual analysis. The GO terms for the genes in (9) are given in (10) and (11).

10) EGR1|[transcription factor activity; regulation of transcription, DNA-dependent; nucleus]

11) AR|[androgen receptor activity; steroid binding; receptor activity; transcription factor activity; transport; sex differentiation; regulation of transcription, DNA-dependent; signal transduction; cell-cell signaling; nucleus]

## 4 SemGen Evaluation and Error Analysis

Before suggesting an application using SemGen output, we discuss the results of error analysis performed on 344 sentences from MEDLINE citations related to six genetic diseases: Alzheimer's disease, Crohn's disease, lung cancer, ovarian cancer, prostate cancer and sickle cell anemia. Out of 442 predications identified by SemGen, 181 were correct, for 41% precision. This is not yet accurate enough to support a production system; however, the majority of the errors are focused in two syntactic areas, and we believe that with further development it is possible to provide output effective for supporting practical applications.

The majority of the errors fall into one of two major syntactic classes, relativization and coordination. A further source of error is the fact that we have not yet addressed interaction relations that involve a process in addition to a gene.

Reduced relative clauses, such as *mediated by Tip60* in (12), are a rich source of argument identification errors.

12) LRPICD dramatically inhibits APP-derived intracellular domain/Fe65 transactivation mediated by Tip60.

SemGen wrongly interpreted this sentence as asserting that LRPICD inhibits Tip60. The rules of the underspecified dependency grammar that identify arguments essentially look to the left and right of a verb for a noun phrase that has been marked as referring to a genetic phenomenon. Arguments are not allowed to be used in more than one predication (unless licensed by coordination or as the head of a relative clause).

A number of phenomena conspire in (12) to wrongly allow *TIP60* to be analyzed as the object of *inhibits*. The actual object, *transactivation*, was not recognized because we have not yet addressed processes as arguments of gene interaction predications. Further, the predication on *transactivation*, with argument *TIP60*, was not interpreted, and hence *TIP60* was available (incorrectly) for the object of *inhibits*. If we had recognized the relative clause in (12), *TIP60* would not have been reused as an argument of *inhibits*, since only heads of relative clauses can be reused.

The underspecified analysis on which SemGen is based is not always effective in identifying verb phrase coordination, as in (13), leading to the incorrect interpretation that WIF1 interacts with SFRP4.

13) WIF1 was down-regulated in 64% of primary prostate cancers, while SFRP4 was up-regulated in 81% of the patients.

A further source of error in this sentence is that *down-regulated* was analyzed by the tagger as a past tense rather than past participle, thus causing the argument identification phase to look for an object to the right of this verb form. A further issue here is that we have not yet addressed truncated passives.

## 5 Using SemGen to Compare OMIM and MEDLINE

SemGen errors notwithstanding, we are investigating possibilities for exploiting automatically extracted gene interaction predications. We discuss an application which compares MEDLINE text to OMIM documents, for specified diseases. LocusLink preferred gene symbols and GO terms are an integral part of this processing. We feel it is instructive to investigate the consequences of this comparison, anticipating results that are effective enough for practical application.

We selected five diseases with a genetic component (Alzheimer's disease, Crohn's disease, lung cancer, prostate cancer, and sickle cell anemia), and retrieved the corresponding OMIM report for each disease, automatically discarding sections such as references, headings, and edit history. We also queried PubMed for each disease and retrieved all MEDLINE citations that were more recent than the corresponding OMIM report. Both OMIM and MEDLINE files were then submitted to SemGen.

For each disease, the MEDLINE file was larger than the corresponding OMIM file, and the categorizer eliminated some parts of each file as not being in the molecular biology domain. Table 1 shows the number of sentences in the original input files and the number processed after the categorizer eliminated sentences not in the molecular biology domain.

| | OMIM Orig. | OMIM Proc. | MEDLINE Orig. | MEDLINE Proc. |
|---|---|---|---|---|
| Alz | 408 | 264 | 1639 | 862 |
| Crohn | 188 | 124 | 4871 | 1236 |
| LungCa | 55 | 34 | 9058 | 2966 |
| ProstCa | 121 | 69 | 6989 | 2964 |
| SCA | 184 | 79 | 4383 | 1057 |

Table 1. Input sentences processed by SemGen

A paragraph in the OMIM file for Alzheimer's disease beginning with the sentence *Alzheimer disease is by far the most common cause of dementia,* for example, was eliminated, while a MEDLINE citation with the title *Semantic decision making in early probable AD: A PET activation study* was removed.

An overview of predication types retrieved by SemGen is given in Table 2 for the files on Alzheimer's disease. Of the gene-disease predications, the majority had predicate ASSOCIATED_WITH (15 from OMIM and 25 from MEDLINE). For gene-gene relations, INTERACT_WITH predominated (3 from OMIM and 12 from MEDLINE).

| **Alzheimer disease** | OMIM | MEDLINE |
|---|---|---|
| Gene-Disease | 16 | 31 |
| Gene-Gene | 3 | 22 |
| Total | 19 | 53 |

Table 2. Gene interaction predication types

We developed a program that compares semantic predications found in MEDLINE abstracts to those found in an OMIM report associated with a particular disease and classifies the comparison between two predications as either an exact match, partial match, or no match. The category of a comparison is determined by examining the argument and predicate fields of the predications. If all three fields match, the comparison is an exact match; if any two fields match it is a partial match. All other cases are considered as no match.

Although fewer than half of the predications extracted by SemGen are likely to be correct, we provide some examples from the files on Alzheimer's disease. (The system retains the document ID's, which are suppressed here for clarity.) Examples of partial matches between gene-disease predications extracted from OMIM and MEDLINE are shown in (14) and (15).

14) OM: APP | ASSOCIATED_WITH | Alzheimer's Disease
ML: CD14 | ASSOCIATED_WITH | Alzheimer's Disease

15) OM: amyloid beta peptide | ASSOCIATED_WITH | Alzheimer's Disease
ML: amyloid beta peptide | ASSOCIATED_WITH | Senile Plaques

Some of the gene-disease predications that only occurred in OMIM are given in (16), and a few of those occurring exclusively in MEDLINE are given in (17).

16) TGFB1 | ASSOCIATED_WITH | Amyloid deposition
PRNP | ASSOCIATED_WITH | Amyloid deposition

Mutation 4 gene | CAUSE | Alzheimer's Disease

17) MOG | ASSOCIATED_WITH | Nervous System Diseases
Acetylcholinesterase | PREDISPOSE | Alzheimer's Disease

In (18) are listed some of the gene-gene interaction predications found in MEDLINE but not in OMIM.

18) LAMR1 | STIMULATE | HTATIP
MAPT|INTERACT_WITH | HSPA8
CD14 | STIMULATE | amyloid peptide

## 6 Using the GO Terms

As noted above, for each gene argument in the predications identified by SemGen, we retrieved from LocusLink the GO terms associated with that gene. We have begun to investigate ways in which these terms might be used to compare genes by looking at the gene-gene interaction predications extracted from MEDLINE that did not occur in OMIM.

To support this work, we developed a program that sorts gene-gene interaction predications by the GO terms of their arguments. For each gene function, the predications in which both arguments share the same function are listed first. These are followed by the predications in which only the first argument has that gene function, and then the predications in which only the second argument has the relevant gene function. A typical output file of this process is shown in (19):

19) RECEPTOR ACTIVITY
---
Both Arguments:
DTR|STIMULATE|EGFR
First Argument:
AR|STIMULATE|TRXR3
EPHB2|STIMULATE|ENO2
Second Argument:
EGR1|INTERACT_WITH|AR
PSMC6|STIMULATE|AR

The three branches of the Gene Ontology provide a uniform system for relating genes by function. The terms in the molecular function and biological process branches are perhaps most useful for this purpose; however, we have begun by considering all three branches (including the cellular component branch). The most effective method of exploiting GO annotations remains a matter of research.

It is important to recognize that GO mapping is not precise; different annotators may make different GO assignments for the same gene. Nevertheless, GO annotations provide considerable potential for relating the molecular functions and biological processes of genes. We consider one of the predications extracted from the MEDLINE file for prostate cancer that did not occur in OMIM:

19) EGR1|INTERACT_WITH|AR

Both genes EGR1 and AR in LocusLink elicit the same human gene set (367 Hs AR; 1026 Hs CDKN1A; 1958 Hs EGR1; 3949 Hs LDLR; 4664 Hs NAB1; 4665 Hs NAB2; 5734 Hs PTGER4; 114034 Hs TOE1). This suggests a high degree of sequence homology and functional similarity. In addition, LocusLink provides the following GO terms for the two genes:

20) EGR1: early growth response 1; LocusID: 1958 Gene Ontology: transcription factor activity; regulation of transcription, DNA-dependent; nucleus

21) AR: androgen receptor (dihydrotestosterone receptor; testicular feminization; spinal and bulb ar muscular atrophy; Kennedy disease) ; LocusID: 367 Gene Ontology: androgen receptor activity; steroid binding; receptor activity; transcription factor activity; transport; sex differentiation; regulation of transcription, DNA-dependent; signal transduction; cell-cell signaling; nucleus

(The GO provides additional, hierarchical information for terms, which we have not yet exploited.)

Thirty percent of the predications examined had some degree of overlap in their GO terms. For example, the terms for EGR1 (transcription factor activity; regulation of transcription, DNA- dependent; and nucleus) are identical to three of the GO terms for the AR gene. This concordance may not be typical of the majority of paired genes in our sample. However, in the case of genes that do not exhibit such complete overlap, concordance might be obtained at higher nodes in the classification scheme.

An alternate approach for assessing distance between GO annotations has been suggested by Lord et al. (2003a, 2003b). They propose a "semantic similarity measure" using ontologies to explore the relationships between genes that may have associated interaction or function. The authors consider the information content of each GO term, defined as the number of times each term, or any child term, occurs.

The fact that any one gene has a number of GO annotations indicates that a particular gene may perform more than one function or its function may be classified under a number of molecular activities. Some of these activities may be part of, i.e. extending to a variable degree down, the same GO structure. For example, for

gene AR, "receptor activity" (GO 4872) partially overlaps with "androgen receptor activity" (GO 4882), as does "steroid binding" (GO 5496) with "transcription factor activity" (GO 3700), and "signal transduction (GO 7165) and "cell-cell signaling (GO 7267). This indicates that in assessing similarity one needs to examine the ontology structure and not rely solely on the GO terms.

While we have no experimental evidence, we would like to speculate about the functional or biological significance indicated by similarity in GO annotation. There are three orthogonal aspects to GO: molecular function, biological process, and cellular component. If two genes map more closely in one of the taxonomies, then their function is necessarily more closely related. The majority of GO terms are in the molecular function taxonomy. It is conceivable that genes that map more closely could be involved in the same cascade or participate in the same genetic regulatory network. There is increasing interest in genetic networks (e.g. www.genome.ad.jp/kegg/ kegg2.html; http://ecocyc.org; http://us.expasy.org/tools/pathways; www.biocarta.com) and combining the ability to search and extract information from the literature with GO mapping could prove effective in elucidating the functional interactions of genes.

## 7 Potential Knowledge Discovery

To determine whether our automatic comparison of MEDLINE to OMIM based on SemGen predications might throw new light on gene-gene interactions, we examined predications found in the MEDLINE file that had no match in the OMIM file. We searched the OMIM reports for information on the genes found in such predications to confirm that they were absent from the OMIM reports. For example, while the OMIM report on colon cancer did not mention BARD1, the SemGen output for MEDLINE had

22) BARD1|INTERACT_WITH|hmsh2

The abstract containing this predication (PMID 11498787) asserts that the BARD1 gene (LocusID 580) interacts with the breast cancer gene BRCA1 as well as with hMSH2, a mismatch repair gene associated with colon cancer. BARD1 shares homology with the two conserved regions of BRCA1 and also interacts with the N-terminal region of BRCA1. Interaction of BARD1 with BRCA1 could be essential for the function of BRCA1 in tumor suppression.

Conversely, disruption of this interaction may possibly contribute to the process of oncogenesis. It has been reported that the BRCA1/BARD1 complex is responsible for many of the tumor suppression activities of BRCA1 (Baer and Ludwig 2002). The gene hMSH2 (LocusID 4436) is one of a number of genes that, when mutated, predisposes to colon cancer type 1. It is the human homolog of the bacterial mismatch repair gene mutS. We hypothesize that the interaction of BARD1 with hMSH2, in a similar fashion to BRCA1, may be necessary for tumor suppression. Disruption of this interaction may increase the likelihood of developing colon cancer. Furthermore, this observation serves to point toward a possible link between BRCA1 and colon cancer.

## 8 Conclusion

We have extended earlier work with SemGen (Rindflesch et al. 2003) and are now able to extract from text, in addition to names of gene and disorders, gene-disorder and gene-gene relations. Although SemGen is not at a stage where it can be used indiscriminately and without selective review and evaluation, it may nevertheless prove useful for reviewers by providing an efficient means of scanning a large number of references and extracting relations involving genes and diseases.

The process of curation and review is time consuming. Given the rate at which new publications are added to the scientific literature, the availability of tools for accelerating the review process would meet a real need. As demonstrated by our pilot study on six disorders, SemGen could prove useful, even at this prototype stage, in extracting relevant information from the literature concerning genes and diseases. Additionally, the ability to scan and extract information from diverse scientific domains could play an important role in identifying new relationships between genes and diseases that would promote hypothesis-generation and advance scientific research. Even with the present limitations, SemGen could assist in making the scientific literature more accessible and reduce the time it takes for researchers to update their knowledge and expertise.

## References


Aronson, A.R. (2001). "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program." In *Proceedings of the AMIA Annual Symposium*, 17-21.

Baer, R., and Ludwig, T. (2002). "The BRCA1/BARD1 heterodimer: a tumor suppressor complex with ubiquitin E3 ligase activity." *Current Opinion in Genetics & Development,* 12, 86-91

Blaschke, C.; Andrade, M.A.; Ouzounis, C.; and Valencia, A. (1999). "Automatic extraction of biological information from scientific text: protein-protein interactions." In *Proceedings of the Seventh International Conference on Intelligent Systems for*

*Molecular Biology*, 60-70.

de Bruijn, B., and Martin, J. (2002). "Getting to the (c)ore of knowledge: mining biomedical literature." *International Journal of Medical Informatics*, 67, 7-18.

Cutting, D.; Kupiec, J.; Pedersen, J.; and Sibun, P. (1992). "A practical part-of-speech tagger." In *Proceedings of the Third Conference on Applied Natural Language Processing*.

Friedman, C., and Hripcsak, G. (1999). "Natural language processing and its future in medicine." *Academic Medicine,* 74 (8),890-5.

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

Gaizauskas, R; Demetriou, G.; Artymiuk, P.J.; and Willett, P. (2003). "Protein Structures and Information Extraction from Biological Texts: The PASTA System." *Bioinformatics*, 19, 135-43.

The Gene Ontology Consortium. (2000). "Gene ontology: tool for the unificaiton of biology." *Nature,* 25, 25-29.

The Gene Ontology Consortium. (2001). "Creating the Gene Ontology Resource: Design and implementation." *Genome Research,* 11,1425-1433.

The Gene Ontology Consortium. (2004). "The Gene Ontology (GO) database and informatics resource." *Nucleic Acids Research*, 32,D258-D261.

Humphrey, S. (1999). "Automatic indexing of documents from journal descriptors: A preliminary investigation." *Journal of the American Society for Information Science*, 50(8), 661-74.

Humphreys, B.L.; Lindberg, D.A.; Schoolman, H.M.; and Barnett, G.O. (1998). "The Unified Medical language System: An informatics research collaboration." *Journal of American Medical Informatics Association*, 5(1), 1-13.

Jenssen, T.K.; Laegreid, A.; Komoroswski, J.; and Hovig, E. (2001). "A literature network of human genes for high-throughput analysis of gene expression." *Nature Genetics*, 28,21-28.

Leroy, G.; Chen, H.; Martinez, J.D. (2003) "A shallow parser based on closed-class words to capture relations in biomedical text." *Journal of Biomedical Informatics*, 36, 145-58 .

Lord, P.W.; Stevens, R.D.; Brass, A.; and Goble, C.A. (2003a). "Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation." *Bioinformatics* 19:1275-1283.

Lord, P.W.; Stevens, R.D.; Brass, A.; and Goble, C.A. (2003b). "Semantic similarity measures as tools for exploring the Gene Ontology." *Pacific Symposium on. Biocomputing*, 601-612.

Masys, D.R.; Welsh, J.B.; Fink, J.L.; Gribskov, M.; Klacansky, I.; and Vorbeil, J. (2001). "Use of keyword hierarchies to interpret gene expression patterns." *Bioinformatics*, 17(4), 319-26.

McCray, A.T.; Srinivasan, S.; and Browne, A.C. (1994). "Lexical methods for managing variation in biomedical terminologies." In *Proceedings of the Annual Symposium on Computer Applications in Medical Care*, 235-9.

Ng, S.K., and Wong, M. (1999). "Toward routine automatic pathway discovery from on-line scientific text abstracts." *Genome Informatics*, 10,104-112.

Online Mendelian Inheritance in Man, OMIM (2000). McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD). WWW URL: http://www.ncbi.nlm.nih.gov/omim/.

Pustejovsky, J.; Castano, J.; Zhang, J.; Kotecki, M.; and Cochran, B. (2002). "Robust relational parsing over biomedical literature: extracting inhibit relations." *Pacific Symposium on Biocomputing*, 362-73.

Rindflesch, T. C., and Fiszman, M. (2003). "The interaction of domain knowledge and linguistic structure in natural language processing: Interpreting hypernymic propositions in biomedical text." *Journal of Biomedical Informatics*, 36(6):462-77.

Rindflesch, T. C.; Libbus, B.; Hristovski, D.; Aronson, A.R.; and Kilicoglu, H. (2003). "Semantic relations asserting the etiology of genetic diseases." In *Proceedings of the AMIA Annual Symposium, 554-8.*

Srinivasan, P., Rindflesch, T.C. (2002). "Exploring Text Mining from MEDLINE." In *Proceedings of the AMIA Annual Symposium*, 722-6.

Stephens, M.; Palakal, M.; Mukhopadhyay, S.; and Raje, R. (2001). "Detecting gene relations form Medline abstracts." In *Proceedings of the Sixth Pacific Symposium on Biocomputing*, 6, 483-96.

Tanabe, L.; Scherf, U.; Smith, L.H.; Lee, J.K.; Hunter, L.; Weinstein, J.N. (1999). "MedMiner: An Internet text- mining tool for biomedical information, with application to gene expression profiling." *BioTechniques*, 27(6),1210-17.

Tanabe, L., Wilbur, W.J. (2002). "Tagging gene and protein names in biomedical text." *Bioinformatics*, 18(8), 1124-32.

Tao, Y-C., and Leibel, R.L. (2002). "Identifying relationships among human genes by systematic analysis of biological literature." *BMC Bioinformatics*, 3,16-25.

Temkin, J. M., and Gilder, M. R. (2003). "Extraction of protein interaction information from unstructured text using a context-free grammar." *Bioinformatics*, 19(16), 2046-53.

Wheeler, D.L.; Church, D.M.; Edgar, R.; Federhen, S.; Helmberg, W.; Madden, T.L.; Pontius, J.U.; Schuler, G.D.; Schriml, L.M.; Sequeira, E.; Suzek, T.O.; Tatusova, T.A.; Wagner, L. (2004). "Database resources of the National Center for Biotechnology Information: update." *Nucleic Acids Research*, 32(1), D35-40.

Yandell, M.D., and Majoros, W.H. (2002) "Genomics and natural language processing." *Nature Reviews Genetics*, 3, 601-610.