Semantic Relations Asserting the Etiology of Genetic Diseases

Thomas C. Rindflesch Ph.D., Bisharah Libbus Ph.D., Dimitar Hristovski Ph.D.,
Alan R. Aronson Ph.D., Halil Kilicoglu M.S.
National Library of Medicine, National Institutes of Health, Department of Health and
Human Services, Bethesda, Maryland 20894

Considerable research is being directed at extracting molecular biology information from text. Particularly challenging in this regard is to identify relations between entities, such as protein-protein interactions or molecular pathways. In this paper we present a natural language processing method for extracting causal relations between genetic phenomena and diseases. After presenting the results of a preliminary evaluation, we suggest the use of a graphical display application for viewing the semantic predications produced by the system.

INTRODUCTION

The ability to identify assertions in the molecular biology literature about the interaction of entities, such as genes and proteins and other phenomena, provides enriched results for applications depending on information about these entities. In research on the genetic basis of disease, the OMIM database [1] contains summaries and curated updates of genes that are associated with human health and disease. Links to the literature are provided by database editors. As a supplement to OMIM resources, automatic methods of discovering associations between genes and diseases have been investigated.

In this paper, we discuss the development of a natural language processing program (called SemGen) to identify and extract semantic propositions on the causal interaction of genes and diseases from MEDLINE citations. We first limit input text to the molecular genetics domain and then use a variety of resources, including the Unified Medical Language System® (UMLS)® Metathesaurus® [2] to automatically identify genetic phenomena and diseases, and then determine relationships asserted between them.

After giving an overview of resources used, we briefly discuss scrutiny of sample text focusing on linguistic patterns commonly used to encode propositions asserting the etiology of genetic disease. We then describe the SemGen mechanism and provide examples of current output as well as the results of a preliminary evaluation. Finally, we explore possibilities for clustering and displaying these results in graphical form.

BACKGROUND

Extraction of semantic relations depends on identifying gene names, proteins, and other genetic phenomena in text; this is challenging and several methods have been proposed [3, 4, 5, for example]. As part of our strategy for this, we use ABGene [6], which employs several statistical and empirical methods.

An array of techniques have been investigated for identifying various relationships in molecular biology text. Methods that emphasize linguistic processing include [7] for molecular pathways, [8] for protein structure, and [9] for protein interactions. Systems directed at gene and disease interactions specifically have been based solely on cooccurrence in the online literature [10], as well as cooccurrence coupled with some linguistic analysis [11].

In order to increase accuracy, SemGen employs a method of identifying MEDLINE citations in the molecular genetics domain before natural language processing begins. This identification is based on journal descriptor indexing (JDI) [12], a statistically-based method for labeled categorization of medical text that draws on human assignment of a small set of descriptors to journal titles in MEDLINE; it does not depend on the MeSH indexing terms assigned to citations.

SemGen is based on SemRep [13], a program being developed to interpret semantic propositions in medical text. SemRep consists of three major components: a) an underspecified syntactic parser that depends on the SPECIALIST Lexicon [14] and the Xerox Part-of-Speech Tagger [15]; b) a component for matching simple noun phrases to concepts in the UMLS Metathesaurus by MetaMap [16]; and c) a mechanism for interpreting semantic relationships based on dependency grammar rules for argument identification and the UMLS Semantic Network for semantic validation of the semantic propositions constructed by the system.

Crucial to the process of semantic interpretation in SemRep is the notion of "indicator rules." These define a correspondence between syntactic elements (such as verbs, nominalizations, and prepositions) and semantic predicates. For example, in the interpretation of *chemotherapy for bladder cancer*, an indicator rule links the preposition *for* to the UMLS Semantic Network predicate TREATS, with arguments 'Therapeutic or Preventive Procedure' and 'Neoplastic Process', which validates the interpretation of this text as "Chemotherapy-TREATS-Bladder Cancer." In this paper we discuss the adaptation of this methodology to recognizing semantic relations on the etiology of genetic diseases.

METHODS

Identifying syntactic patterns and semantic structure

Before constructing SemGen, we determined common indicator rules for semantic relationships on the interaction of genes and diseases by scrutinizing a training sample of characteristic text. Twenty verbs (and their nominalizations) were determined to encode a relation between a genetic phenomenon and a disorder. Two prepositions, *in* and *for* (cf. [17]), were also noted during this process.

The twenty indicators found during this analysis were involved in predications asserting some sort of a causal relation between a genetic phenomenon and a disorder, in particular: CAUSE (indicated by *cause*, *determine*, and *underlie*, for example), PREDISPOSE (*predispose, lead to, susceptibility*), and ASSOCIATED_WITH (*associated with, involve, related, in*).

We considered these three semantic relations to be children of the more general relation ETIOLOGY. Further, the three more specific relations are in a strength hierarchy, which is from strongest to weakest: CAUSE → PREDISPOSE → ASSOCIATED_WITH.

We submitted the training sample to MetaMap; Metathesaurus concepts and their semantic types found in the text formed the basis of generalizations about the semantic characteristics of the arguments in the relevant predications. In addition to the semantic type 'Gene or Genome', we stipulated other semantic types, including 'Nucleotide Sequence', as allowable subjects of an etiology relation. Additional genetic processes and entities, such as mutations, polymorphisms, and chromosomes, were included in the definition of the subject of any of the etiology relations we address. Similarly, the semantic class for the object of these relations includes the semantic type 'Disease or Syndrome' as well as additional semantic types such as 'Neoplastic Process' and 'Congenital Abnormality'.

This semantic framework, including the three predicates noted in a strength hierarchy and as children of ETIOLOGY, with subject defined as the semantic class <genphenom> and object as <disorder> serves as the underpinning for SemGen.

Identifying etiology relationships in text

As noted earlier, SemGen is a modification and enhancement of SemRep. It is enhanced first with a labeled categorizer and secondly with mechanisms for identifying genetic phenomena. These are taken from Edgar [18,19] and Arbiter [20] as well as ABGene [6]. The identification of disorder concepts depends on MetaMap and the UMLS Metathesaurus. Once the referential vocabulary has been identified, existing SemRep processing interprets the relational vocabulary.

SemGen begins by sending a MEDLINE citation (title and abstract) to the labeled categorizer. Text meeting the criterion of a rule designed to recognize content in the molecular genetics domain is passed on for further processing, while anything else is ignored. The title and abstract of filtered citations are then sent to ABGene and genes found anywhere in the text are returned in a list that contributes to the identification of genetic phenomena during processing of the referential vocabulary.

In linguistic processing, the initial phase is identical to SemRep. After lexical look-up and tagging, an underspecified parse serves as the basis for the identification of the referential vocabulary. For example, the syntactic structure for (1) is represented schematically in (2).

(1) Deletions of the INK4A gene occur in malignant peripheral nerve sheath tumors but not in neurofibromas.

(2) [deletions] [of the INK4A gene] [occur] [in malignant peripheral nerve sheath tumors] [but] [not] [in neurofibromas]

Each simple noun phrase from the underspecified parse is subjected to three steps. First, MetaMap attempts to identify concept in the Metathesaurus. If the corresponding semantic type belongs to a set of semantic types for genetic phenomena (such as 'Gene or Genome') that noun phrase is considered to refer to a genetic phenomenon. Second, the text tokens in the phrase are compared to the list of gene names earlier received from ABGene. A match qualifies the noun phrase as referring to a genetic phenomenon. In the third step, words in a noun phrase not having met one of the first two criteria are matched against a small list of characteristic words for genetic phenomena, such as *codon*, *exon*, *deletion*, etc. Finally, contiguous genetic noun phrases are coalesced into a single macro noun phrase, which is considered to be a potential subject in a semantic relationship on the genetic etiology of disease.

The application of this processing to (2) identifies *deletions* (characteristic word) and *of the INK4 gene* (ABGene) as genetic phenomena. These are coalesed into *deletions of the INK4A gene.*

Identification of the referential vocabulary for disorders depends only on MetaMap and the Metathesaurus. Noun phrases matching a concept having a semantic type for disorders are marked as potential objects in the semantic predications constructed during the next phase of processing. In (2), disorder concepts “Malignant Peripheral Nerve Sheath Tumors (semantic type ‘Neoplastic Process’) and “Neurofibroma” (‘Neoplastic Process’) are found.

During the next phase of SemGen processing, interpretation of the relational vocabulary, semantic predications asserting an etiological relationship between genetic phenomena and disorders are constructed. The application of indicator rules identifies the predicate in this relationship. In (2) above, the preposition *in* encodes the semantic predicate ASSOCIATED_WITH.

The dependency grammar rules that identify arguments are satisfied if the subject is to the left of the indicator and the object is to the right. Further, a noun phrase cannot be reused in the construction of a predication, without license. Semantic validation for the construction of these predications is conferred by the constraint that the subject must be a genetic phenomenon and the object must be a disorder. These constraints can only be met by interpreting the predications in (3) and (4) for (2), in which coordination licenses reuse of the subject.

(3) deletion ink4a gene
|ASSOCIATED_WITH|Malignant Peripheral
Nerve Sheath Tumors

(4) deletion ink4a gene
|NEG_ASSOCIATED_WITH|Neurofibroma

We performed a preliminary evaluation based on a post hoc sample of 1,000 sentences extracted at random from SemGen output. The evaluation is limited in that the judge was one of the members of the team (BL) and only false positives were marked.

RESULTS

We issued a PubMed search with the query “p53 OR mdm2,” which retrieved 27,485 citations; these were then processed by SemGen. Twenty-two percent (6,111) of the citations that matched the query were eliminated by the labeled catgorizer as not being in the molecular genetics domain. We did not check for false positives, that is citations that were processed, but did not pertain to molecular genetics. Random scrutiny of the citations that were eliminated indicated that these were often about the biological mechanisms of the relevant genes, rather than their relationship to disease.

SemGen extracted 24,716 semantic relationships from the 21,374 citations that passed the JDI labeled categorizer. An overview of the distribution of the relationships is given in (5), where the relative frequency of the three positive relationships reflects the strength hierarchy mentioned above.

(5) 22861 ASSOCIATED_WITH
845 PREDISPOSE
482 NEG_ASSOCIATED_WITH
479 CAUSE
25 NEG_PREDISPOSE
24 NEG_CAUSE

Sentences for evaluation were also extracted from this output. SemGen identified 1,124 relationships in the 1,000 sentences scrutinized. Of these 271 were marked as wrong, leaving 853 correct. Precision was thus 76%.

DISCUSSION

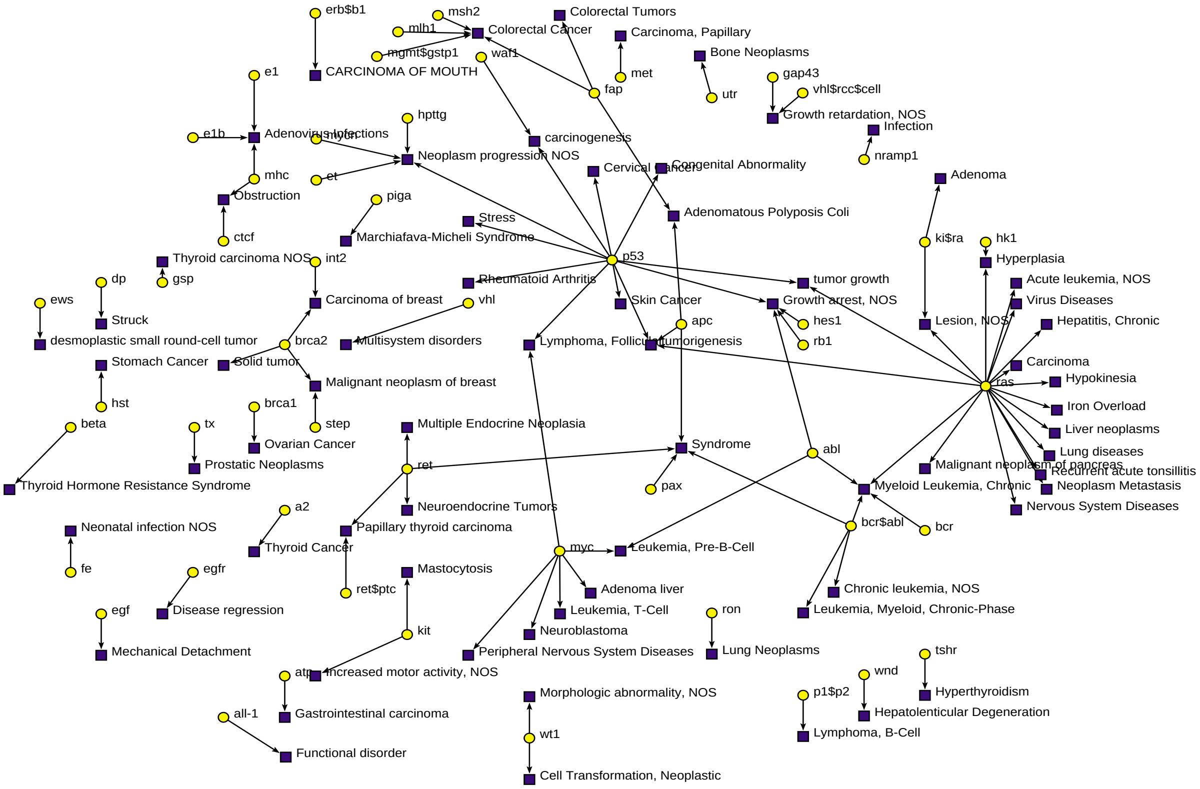
Error analysis revealed that of the 271 incorrect etiology predications, in only 19 was the predicate wrong. In all such cases, due to inadequate negation processing in SemGen, the predicate was stated positively, when it had been asserted negatively in the text.

In the remaining 252 errors, one (or both) of the arguments had been inaccurately identified. There were 105 mistakes in the genetic phenomenon argument and 147 disorders were wrong.

Most of the errors in disease recognition were due to word sense ambiguity as represented in the Metathesaurus. For example, the concept “Recurrent acute tonsillitis” has the synonym “RAT” in the Metathesaurus, and SemGen erroneously mapped text *rat* to that concept. The spurious occurrence of this disease in the analysis then caused an erroneous predication to be generated.

The majority of the errors in disease recognition involved a failure to identify the most specific concept asserted in the text. For example, text *human myeloid leukemogenesis* mapped only to the general concept “leukemogenesis,” since the specific term does not occur in the Metathesaurus.

A similar phenomenon characterized errors in gene recognition. Whereas errors in disease recognition were largely due to the Metathesaurus, errors in recognizing specific genes were most often due to errors in syntactic processing. For example from text *the Ki-ras proto-oncogene* ABGene recognized *Ki-ras* as a gene name, but subsequent SemGen processing only recognized *oncogene* as a term available to be interpreted as the subject of an etiology predication. The emphasis in future work will be to address both word sense ambiguity in this domain and to enhance recognition of the more specific concept.



We extracted just the CAUSE relations from the SemGen output returned from the PubMed query on the oncogenes p53 and mdm2, and submitted them to Pajek [21] (available at http://vlado.fmf.uni-lj.si/pub/networks/pajek/) for visualization of the network representing these relations in the citations processed. In the diagram above, genes (circles) and diseases (squares) are represented by vertices; since only one relation is represented, arcs are not labeled.

Visual representation such as this facilitates the identification of genes related to more then one disease and diseases influenced by more than one gene. An extension of this methodology is to cluster networks such as this in order to provide a view in which similar genes are grouped together based on the diseases they influence.

Finally, SemGen may be useful for literature-based discovery [22,23]. Current systems are based on cooccurrence of concepts in MEDLINE citations. The cooccurrence of a particular gene and disease may indicate that a discovery has not been made. Using SemGen, however, the existence of a PREDISPOSE or ASSOCIATED relation between a gene and a disease in the literature might still be interesting. Such relations might indicate that a CAUSE relation between these entities is close to being discovered—the ultimate goal of the system.

CONCLUSION

We have presented a natural language processing system for extracting semantic relations expressing the genetic basis of disease from MEDLINE citations. The method discussed is based on the enhancement and integration of several existing resources, including the UMLS Metathesaurus. The results of a preliminary evaluation are encouraging. We also suggest the use of a graphical display application for clustering and viewing the semantic predications extracted from MEDLINE citations.

References

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

2. Humphreys BL, Lindberg DAB, Schoolman HM, and Barnett GO. The Unified Medical language System: An informatics research collaboration. JAMIA 1998:5(1):1-13.

3. Fukuda K, Tamura A, Tsunoda T, Takagi T. Toward information extraction: identifying protein names from biological papers. Pac Symp Biocomput. 1998;:707-18.

4. Krauthammer M, Rzhetsky A, Morozov P, Friedman C. Using BLAST for identifying gene and protein names in journal articles. Gene. 2000 Dec 23;259(1-2):245-52.

5. Majoros WH, Subramanian GM, Yandell MD. Identification of key concepts in biomedical literature using a modified Markov heuristic. Bioinformatics. 2003 Feb;19(3):402-7.

6. Tanabe L, Wilbur WJ. Tagging gene and protein names in biomedical text. Bioinformatics. 2002;18(8):1124-32.

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

8. Gaizauskas R, Demetriou G, Artymiuk PJ, Willett P. Protein Structures and Information Extraction from Biological Texts: The PASTA System. Bioinformatics. 2003 Jan;19(1):135-43.

9. Pustejovsky J, Castano J, Zhang J, Kotecki M, Cochran B. Robust relational parsing over biomedical literature: extracting inhibit relations. Pac Symp Biocomput. 2002;:362-73.

10. Adamic LA, Wilkinson D, Huberman BA, Adar E. A literature based method for identifying gene-disease connections. IEEE Computer Society Bioinformatics Conference. 2002:;109-117.

11. Craven M, Kumlien J. Constructing biological knowledge bases by extracting information from text sources. Proc Int Conf Intell Syst Mol Biol. 1999;:77-86.

12. Humphrey S. Automatic indexing of documents from journal descriptors: A preliminary investigation. JASIS 1999;50(8):661-674.

13. Srinivasan P, Rindflesch TC. Exploring Text Mining from MEDLINE. Proc AMIA Symp. 2002;:722-6.

14. McCray AT, Srinivasan S and Browne AC. Lexical methods for managing variation in biomedical terminologies. Proc SCAMC, 1994, 235-9.

15. Cutting D, Kupiec J, Pedersen J and Sibun P. A practical part-of-speech tagger. Proceedings of the Third Conference on Applied Natural Language Processing, 1992

16. Aronson A. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001;:17-21.

17. Leroy G, Chen H. Filling preposition-based templates to capture information from medical abstracts. Pac Symp Biocomput. 2002;:350-61.

18. Rindflesch TC, Tanabe L, Weinstein JW, Hunter L. EDGAR: extraction of drugs, genes and relations from the biomedical literature. Pac Symp Biocomput. 2000;:517-28.

19. Libbus B, Rindflesch TC. NLP-Based Information Extraction for Managing the Molecular Biology Literature. Proc AMIA Symp. 2002;:445-9.

20. Rindflesch TC, Rajan J, Hunter L. Extracting molecular binding relationships from biomedical text. Appl. Nat. Lang. Proc. 2000;:188-95.

21. Batagelj V, Mrvar A, Zaversnik M. Partitioning approach to visualization of large graphs. Lect. Notes Comp. Sc. 1999;1731:90-97.

22. Swanson DR, Smalheiser. An interactive discovery system for finding complementary literatures: a stimulus to scientific discovery. Artif. Intell. 1997;91:183-202.

23. Hristovski D, Stare J, Peterlin B, Dzeroski S. Supporting discovery in medicine by association rule mining in Medline and UMLS. Medinfo. 2001;10(Pt 2):1344-8.