NLP-Based Information Extraction for Managing the Molecular Biology Literature

Bisharah Libbus[†] Ph.D., Thomas C. Rindflesch[‡] Ph.D.,
[†]University of North Carolina, Chapel Hill, North Carolina
[‡]National Library of Medicine, Bethesda, Maryland

We present research aimed at devising a tool for using natural language processing to identify and extract biomedical information from text for the purpose of assisting researchers in molecular biology manage large amounts of information. A pilot project based on the molecular genetics of diabetes demonstrates our ability to explore the interaction of genomic phenomena and clinical findings. We suggest the cooperation of this extracted information with systems for clustering text and constructing labeled networks of data.

INTRODUCTION

The availability of a vast amount of literature online (particularly in the MEDLINE database) provides the biomedical scientist with instant and practically total accessibility. The researcher, however, is then faced with the daunting task of identifying a manageable subset of these citations relevant to current research requirements. The astounding increase in genomic databases in recent years is a further demonstration of a similar change in the dynamic of data accessibility. While ever-increasing amounts of data are publicly available and instantly accessible, our ability to screen, relate, and interpret data remains less than optimal.

The goal of the research presented here is to construct a tool for identifying and extracting biomedical information from text in order to address this need. Reseach in textual analysis supporting molecular biology (see [1] for an overview) is focused in three broad areas: a) extracting terms and relations from text, b) manipulating terms and relations found in text, and c) exploiting the indexing terms associated with abstracts in MEDLINE citations.

Due to the inherent complexity of language a challenging area of research focuses on the accurate identification of genomic structures, proteins, and other phenonena in the scientific literature ([2], [3], for example). Related work (such as [4], [5], [6], [7]) uses natural language processing (NLP) techniques to retrieve the relationships in which these entities are involved.

Rather than concentrate on the discovery of terms and relations in text, other studies take this information as input and then manipulate it in order to provide the researcher with structured output, which might take the form of a network of terms (perhaps with the relations in the network labeled), or clusters of documents or sentences. In the system described in [8], for example, output is a labeled graph of relations among the input genes. Jenssen and Vinterbo [9] produce a cluster of documents discussing relationships among a set of input genes. The MedMiner system [10] identifies sentences from MEDLINE citations with user-specified terms and relations highlighted.

The National Library of Medicine adds significant value to citations in MEDLINE, and this information can profitably be used in conjunction with data extracted from text automatically. In addition to the traditional MeSH indexing, of particular interest to researchers in molecular biology are the chemical substances and links to molecular sequence data bases. Masys et al. 2001 [11] describe the exploitation of this information in interpreting output derived from high-density microarray technology. Jenssen et al. 2001 [12] discuss a system that takes advantage of the information assigned to indexing fields as well as that extracted from the text of citations in order to construct a gene-to-gene co-citation network for a large number of human genes.

Our ultimate aim is to construct a multi-faceted research tool that takes advantage of the insights gained in previous work to provide as much information as possible to the molecular biology scientist. Ideally such a tool would exploit terms and relations identified automatically in text by both statistical and symbolic NLP methods, in addition to information supplied by NLM indexers. Output would consist of document clusters as well as structured information based on terms and relations found in the text. This information could be used to help navigate through large amounts of both textual and genomic data, providing links and structured output, which could help the user understand current research as well as stimulate scientific discovery [13]. For the remainder of the paper, we discuss a pilot project to demonstrate the usefulness of this endeavor, focusing on the genetic basis of diabetes in general and of non-insulin-dependent diabetes mellitus (NIDDM) in particular.

We selected diabetes as a representative of complex, i.e. multigenic, human traits. Even though the genetic predisposition to the disease is well recognized, and has been for some time, the identification of the specific genes involved still eludes investigators ([14], [15], [16], [17]). For this reason, a thorough and robust literature search could possibly uncover new relationships between gene and disease, and point to alternative associations that could suggest new experimental approaches and treatments.

Before looking at specific methods for automatically processing documents in support of this project, we addressed the general question of the rhetorical structure of research reports in this domain, identifying major components in such research. Scrutiny of the literature allowed us to identify crucial categories in three broad areas of research on the genetic basis of disease: genomic phenomena relating to genes and mutations; clinical concerns such as disease and associated findings; and research variables, including genetic research techniques, subjects, and conclusions.

As a pilot project, we decided to concentrate on genomic information and clinical considerations, limiting ourselves to the following variables: Disease, Findings, Genes, Alleles, Mutations, Variants, Polymorphisms, Genotypes, and Chromosomes. We used NLP techniques to extract values for these variables from the title and abstract of MEDLINE citations. An example of the current output from the system is given in (2) for the abstract whose title appears in (1).

(1) Title: Genes and environment in type 2 diabetes and atherosclerosis in aboriginal Canadians.

(2) Abstract ID: 21185374
Target Disease: Diabetes Mellitus, Non-Insulin-Dependent; Diabetes
Other Diseases: Atherosclerosis; Coronary Artery Disease
Findings: CANADIAN; Obesity; Related; Presence
Genes: hnf1a s319; s319; hnf1a
Alleles: none
Mutations: hnf1a
Variants: none
Polymorphisms: none
Genotypes: none
Chromosomes: none

Although word-sense ambiguity led to spurious findings such as Related and Presence, the identification of genomic phenomena and clinical findings associated with NIDDM in this abstract are encouraging. Such results appear to indicate that further development of this research project is warranted. We would like to report on our progress to date and outline future enhancements that are intended to expand the scope of this research effort**,** improve the accuracy of the identification of the values addressed, and extract relationships between disease and genes and among the identified genetic components.

METHODS

In order to identify genes and related entities we modified an existing Prolog program [18], which was originally intended to support research in molecular phamacology for cancer and which identifies genes, cells, and drugs in text. In addition, we had access to a more effective program for identifying gene names in text [3]. For the identification of diseases and findings we used MetaMap [19], a program that maps text to concepts in the Unified Medical Language System (UMLS) Metathesaurus.

The system pursues two paths in parallel to identify values for the variables discussed above. One path is based on [18] and relies on the SPECIALIST Lexicon and a stochastic tagger [20] to produce an underspecified syntactic parse structure [4] in which simple noun phrases are identified, as shown schematically in (3).

(3) ... suggested [evidence]$_{NP}$ for [linkage] $_{NP}$ of [type 2 diabetes] $_{NP}$ and [human chromosome 20q13] $_{NP}$, [a region] $_{NP}$ ...

This structure then serves as the basis for two techniques to identify the information of interest. The first calls on MetaMap to map the phrases provided by the parser to concepts in the UMLS Metathesaurus [21]. Each concept in the Metathesaurus is categorized with one or more semantic types, such as ‘Disease or Syndrome’ ‘Finding’, or ‘Amino Acid, Peptide, or Protein’, which allow the program to identify diseases and findings, as well as some genomic phenomena. The concepts Glucokinase and Hepatocyte Nuclear Factor, for example, appear in the Metathesaurus with semantic type ‘Amino Acid, Peptide, or Protein’, and hence are identified by our program as genes when encountered in text. Further, in example (3), MetaMap determines that the noun phrase *type 2 diabetes* corresponds to the Metathesaurus concept Diabetes Mellitus, Non-Insulin-Dependent, with semantic type ‘Disease or Syndrome’.

We also use cue words such as *mutation, gene, chromosome*, etc. occurring in a noun phrase to identify values for the variables in this project. The list of cue words for this purpose was compiled during the earlier analysis of the literature on molecular genetics and is subject to emendation and enhancement. Since the noun phrase *human chromosome 20q13* in (3) above contains the cue word *chromosome*, *20q13* is determined to be a value for that variable. Each sentence in the title and abstract of a MEDLINE citation is processed separately by this path (both MetaMap and cue word identification). Subsequently, all information for each abstract is amalgamated [22].

The other path is described in [3] in more detail. It appeals to several statistical and empirical methods to identify gene and protein names. An example of the output from this processing, in which all text tokens are labeled, is given in (4); the genes *PPAR gamma*, *A12A/c1431c*, and *CC A12A/t1431t* have been identified. The information provided by this processing is added to that gleaned by path one, for each citation.

*(4) DNA/NN sequencing/VBG revealed/VBN no/DT additional/JJ mutations/NNS in/IN the/DT coding/JJ region/NN of/IN the/DT PPAR/**MULTIGENE** gamma/**MULTIGENE** gene/**MULTIGENE** in/IN genotypes/NNS A12A/c1431c/**GENE** or/CC A12A/t1431t/**GENE** ./.*

It is important to note that we identified values for variables such as gene, allele, disease, and findings. The investigator need not start with a predetermined list of genes or diseases. This provides the advantage of an open-ended process that facilitates knowledge discovery and information extraction.

RESULTS

Preliminary results were produced by running the program on a sample of 1,075 MEDLINE citations dealing with the molecular genetics of diabetes. Output is provided in two formats, one suitable for humans (similar to that shown in (2) above, with the entire text of the abstract given) and another, machine-readable, version that forms the basis of further computational processing.

Although the output produced by our system contains errors due to limitations on current NLP techniques, it nonetheless provides a useful basis for investigations in the genetic basis of disease. We used a data base as well as Unix commands to generate distributional and cooccurrence information. This processing could benefit from statistical techniques such as those discussed above ([8] and [9], for example). This basic data affords valuable clues to a researcher initially confronting a large amount of textual data. For example, a list of the unique genes that are reported to be associated with NIDDM and the number of citations where they are mentioned provides an overview of the corpus. A partial listing is given in (5).

(5) 33 glucokinase
29 insulin
22 hnf
21 type ii
19 hnf-1alpha
18 insulin receptor
17 mody3
17 hepatocyte nuclear factor
15 mody1
15 glycogen synthase
13 ace
12 insulin receptor substrate-1

A list of all gene name tokens that cooccur with NIDDM along with the citation unique identifiers can be sorted by the identifiers or by the gene names. From the first list, one can retrieve a list of genes mentioned in any citation of interest. The citation in (6), for example, mentions an array of genes associated with NIDDM.

(6) 21232235|3-kinase
21232235|glut4
21232235|glycogen synthase
21232235|hexokinase ii
21232235|insulin receptor
21232235|insulin receptor substrate-1
21232235|mrna
21232235|p110alphapi3k
21232235|p110betapi3k
21232235|p85alpha
21232235|p85alphapi3k
21232235|phosphatidylinositol
21232235|pi3k
21232235|srebp-1c
21232235|srebp-1c mrna

Similar information can be generated for the cooccurrence of more than one gene with a particular disease or for the interaction of genes and findings with NIDDM. For example, one may be interested in determining whether information about two particular genes is included in the same citation. A search for two genes, the hepatocyte nuclear factor (hnf) and glucokinase, generated the following (partial) list.

(7) 20053748|gene|hnf-1alpha|gene|glucokinase
20053748|gene|hnf-1alpha|gene|glucokinase genes
20053748|gene|hnf-4alpha|gene|glucokinase
20053748|gene|hnf-4alpha|gene|glucokinase genes
20136394|gene|hnf|gene|enzyme glucokinase
20136394|gene|hnf-1 alpha|gene|enzyme glcokinase
20136394|gene|hnf-1 beta|gene|enzyme glucokinase
20136394|gene|hnf-3 beta|gene|enzyme glucokinase

It is also possible to search for clinical findings that are associated with a disease. Our output indicates that in addition to obesity, findings such as hypertension, lymphopenia, and deafness frequently occur in citations discussing NIDDM. The co-occurrence of particular genes with deafness in NIDDM patients may be of interest, and a search yielded the following (truncated) list of citations. Data in these abstracts indicate that an A-to-G mutation in the mitochondrial gene for tRNA(Leu) has been associated with hearing loss.

(8) gene|20046384|d20s197
gene|20046384|gck2
gene|20046384|glucokinase gene markers gck1
gene|21185209|slc19a2
gene|21185209|thiamine transporter
gene|94252465|mtdna sequence
gene|95308350|chinese family
gene|96177277|glucose-regulated insulin
gene|96423002|leukocyte dna

DISCUSSION

In addition to taking an “inventory” of the various parameters associated with diabetes, we were able to extract co-occurrence relationships between any two or more terms of choice. Citations that deal with two particular genes, or with a gene and a disease entity, or with a gene and a particular finding, could be listed by constructing queries. More complex searches that involve three variables could be as easily set up. For example, we searched the database for citations that dealt with NIDDM, genes, and deafness. One can, similarly, set up searches involving any number of variables.

The MEDLINE citation contains a number of fields that could be mined for information in addition to that contained in the title and the abstract, for example, information about nucleic acids, proteins, or chemicals. The unique identifiers for these could be retrieved and incorporated in a search query or used to retrieve the sequence of the nucleic acid from GenBank.

The National Center for Biotechnology Information at NLM provides a number of databases [23] that could profitably be used in cooperation with output from our system to provide extensive molecular biology information on the genes and loci of interest. For example, by entering a query for one of the genes in our study, hnf1, in the Online Mendelian Inheritance in Man (OMIM) database, one gets a list of genes and associated human diseases. Under TRANSCRIPTION FACTOR 1 (TCF1) (#142410), the standard name for this gene, there are alternative gene names and their chromosomal band positions, reviews and summaries of several publications, references and links to articles. The LocusLink database further provides a list of gene loci, with their description, chromosomal location, and the organism in which they were studied.

While the approach we are pursuing has the benefit of facilitating searches, a particular advantage is the potential to uncover new relationships that may have evaded detection ([13], [24], [25], [26]). As science becomes more interdisciplinary, the need for complementary approaches and teamwork becomes more recognized. At the same time it is possible that workers may be unaware of the work of others who are addressing the same disease entity or biological process but who use different tools and are active in different disciplines. Relationships that may have so far eluded detection could thus be identified as disparate but complementary citations are captured and indexed.

CONCLUSION

Even though text presents well-known challenges and variables that are not easily resolved by automated analysis, the power of NLP can support the process of knowledge extraction. We have used a variety of techniques and search algorithms to survey a field, in this case the genetic control of diabetes, without prior knowledge or selection of query terms. By specifying the variables of interest we were able to extract specific information and comprehensive listing for such phenomena as genes, mutations, polymorphisms, and disease findings. In an operational system, distributional and cooccurrence information would be generated for all variables and made available to techniques for displaying clusters of citations and networks of the data and providing links between the two.

Acknowledgments

The first author is supported by an NLM Medical Informatics Training Grant. The proje ct was begun during an NLM rotation for NLM medical informatics trainees. We are grateful to Jim Mork for help with database management and to Lorraine Tanabe and John Wilbur for making their program for identifying genes and proteins available for this project.

References

1. Andrade MA, Bork P. Automated extraction of information in molecular biology. FEBS Letters 2000;476:12-7.

2. Fukuda K, Tsunoda T, Tamura A, Takagi T. Toward information extraction: Identifying

protein names from biological papers. Pac. Symp. Biocomput., 1998, 707-18.

3. Tanabe L, Wilbur WJ. Tagging gene and protein names in biomedical text. Bioinformatics, in press.

4. Rindflesch TC, Rajan JV, Hunter L. Extracting molecular binding relationships from biomedical text. Appl. Nat. Lang. Process., 2000:188-95.

5. Hahn U, Romacher M, Schulz S. Creating knowledge repositories from biomedical reports the MEDSYNDIKATE text mining system. Pac. Symp. Biocomput., 2002, 338-49.

6. 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;1(1):1-9.

7. Thomas J, Milward D, Ouzounis C, Pulman S, Carroll M. Automatic extraction of protein interactions from scientific abstracts. Pac Symp. on Biocomp., 2000: 538-49.

8. Stephens M, Palakal M, Mukhopadhyay S, et al. Detecting Gene Relations from MEDLINE Abstracts. Pac. Symp. Biocomput., 2001:483-96.

9. Jenssen TK, Vinterbo S. A set-covering approach to specific search for literature about human genes. Proc. AMIA Symp., 2000:384-8.

10. Tanabe L, Scherf U, Smith LH, Lee JK, Hunter L, Weinstein JN. MedMiner: An Internet text-mining tool for biomedical information, with application to gene expression profiling. BioTechniques, 1999;27(6):1210- 17.

11. Masys DR, Welsh JB, Lynn Fink J, et al. Use of keyword hierarchies to interpret gene expression patterns. Bioinformatics 2001;17(4):319-26.

12. Jenssen TK, Laegreid A, Komorowski J, Hovig E. A literature network of human genes for high-throughput analysis of gene expression. Nat Genet. 2001;28(1):21-8.

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

14. Elbein SC, Hoffman MD, Teng K, et al. A genome-wide search for type 2 diabetes susceptibility genes in Utah Caucasians. Diabetes, 1999; 48:1176-82.

15. Ehm MG, Karnoub MC, Sakul, H, et al. Genomewide search for type 2 diabetes susceptibility genes in four American populations. Am J Hum Genet, 2000; 66:1871-81.

16. Watanabe RM, Ghosh S, Langefeld CD, et al. The Finland-United States investigation of non-insulin-dependent Diabetes Mellitus Genetics (FUSION) Study. II. An sutosomal genome scan for diabetes-related quantitative-trait loci. Am J Hum Genet, 2000; 67:1186-1200.

17. Almind K, Doria A, Kahn CR. Putting the genes for type II diabetes on the map. Nature Med, 2001; 7:277-9.

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

19. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. Proc. AMIA Symp., 2001:17-21.

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

21. Humphreys B. L., Lindberg D. A. B., Schoolman H. M., and Barnett G. O. (1998) The Unified Medical language System: An informatics research collaboration. JAMIA 1998;5(1):1-13.

22. Ding J, Berleant D, Nettleton D, Wurtele E. Mining MEDLINE: Abstracts, Sentences, or Phrases? Pac. Symp. Biocomput., 2002, 326-37.

23. Wheeler DL, Church DM, Lash AE, Leipe DD, Madden TL, Pontius JU, Schuler GD, Schriml LM, Tatusova TA, Wagner L, Rapp BA. Database resources of the National Center for Biotechnology Information: 2002 update. Nucleic Acids Res. 2002;30(1):13-6.

24. Hristovski D, Stare J, Peterlin B, Dzeroski S. Supporting discovery in medicine by association rule mining in Medline and UMLS. Medinfo., 2001:1344-8.

25. Srinivasan P. MeSHmap: A text mining tool for MEDLINE. Proc. AMIA Symp., 2001:642-6.

26. Weeber M, Klien H, Aronson AR, et al. Text-based discovery in biomedicine: the architecture of the DAD-system. Proc. AMIA Symp., 2000:903-7