Semantic MEDLINE: A Web Application for Managing the Results of PubMed Searches

Halil Kilicoglu[1,2], Marcelo Fiszman[1], Alejandro Rodriguez[1], Dongwook Shin[1], Anna M. Ripple[1] and Thomas C. Rindflesch[1]

[1] National Library of Medicine, Bethesda, MD, 20892, USA
[2] Concordia University, Department of Computer Science and Software Engineering, Montreal, QC, H3G 1M8, Canada

Abstract

We describe Semantic MEDLINE, a Web application that manages the results of PubMed searches by summarizing and visualizing semantic predications extracted from MEDLINE citations and linking them to several structured resources to provide an integrated environment. To demonstrate its utility, we present a scenario in which we use Semantic MEDLINE to gain insights into relaxin, a hormone whose function in humans has not been fully elucidated. We propose Semantic MEDLINE as an enabling information resource and exploration tool for biomedical scientists, health care professionals, and consumers. (For access, send e-mail to trindflesch@mail.nih.gov).

1 Introduction

Traditional information retrieval tools often challenge users with the large number of items returned. In the biomedical domain, PubMed provides access to over 17 million citations from some 5000 journals in the MEDLINE database. Sophisticated knowledge management applications are needed to help the user exploit this massive amount of text. Similarly, the amount of structured online health-related information, including biomedical vocabularies, ontologies, clinical and molecular biology knowledge bases, and model organism annotation databases, is growing at a rate that outpaces the development of effective access applications.

Linking the biomedical literature and structured resources presents new opportunities in user-driven text mining and knowledge discovery as well as automatic curation of biomedical resources. We are developing a Web application, called Semantic MEDLINE, which integrates PubMed with natural language processing, automatic summarization, visualization, and interconnections among multiple sources of relevant biomedical information. The system is intended to help health care professionals and consumers keep abreast of current research as well as assist researchers in mining the literature to generate hypotheses. In this paper, we first describe Semantic MEDLINE and its implementation and then discuss a scenario using the tool to elucidate the peptide hormone relaxin.

2 Related Work

Natural language processing often underpins applications in biomedicine, and some systems extract relations from text (Blaschke *et al.*, 1999; Friedman *et al*., 2001; Leroy *et al*., 2003; Lussier *et al*., 2006; Rindflesch and Fiszman, 2003). Others focus on using the information extracted; examples include automatic summarization (McKeown *et al.*, 2001, Fiszman *et al*., 2004a), question answering (Demner-Fushman and Lin, 2007; Jacquemart and Zweigenbaum, 2003; Sable *et al*., 2005; Sneiderman *et al.*, 2007; Wedgwood, 2005), and literature-based knowledge discovery (Ahlers *et al*., 2007b; Hristovski *et al.,* 2006; Srinivasan and Libbus, 2004; Swanson, 1986).

Several recent systems visualize the information extracted. Ali Baba (Plake *et al*., 2006) relies on concepts co-occurring in documents to represent text as a graph of interrelated relationships. Based on co-occurrences of genes in MEDLINE abstracts, Jensen *et al.* (2001) construct networks of genes found relevant in gene expression data analysis. The Telemakus project (Fuller *et al.*, 2004) is based on relationships identified by hand and is meant to enable knowledge discovery through interactive visual maps of linked concepts among documents. The LitMiner system (Feldman *et al.*, 2003) represents several gene-related relations extracted with a type of underspecified natural language processing in a graph. Finally, the PGViewer tool (Tao *et al.*, 2005) visualizes genomic information across both structured and textual databases. Integrating

the biomedical literature with external databases and ontologies has also been explored: GoPubMed (Doms and Schroeder, 2005) and CiteXplore (http://www.ebi.ac.uk/citexplore).

3 Background

At the core of Semantic MEDLINE are two existing tools: SemRep (Rindflesch and Fiszman, 2003), which extracts semantic predications (subject-predicate-object triples) from text, and an automatic summarizer (Fiszman *et al.,* 2004a).

3.1 SemRep

SemRep was developed for the biomedical research literature and uses domain knowledge provided by the Unified Medical Language System (UMLS) (Lindberg *et al.*, 1993). It represents textual content with semantic predications consisting of UMLS Metathesaurus concepts as arguments and UMLS Semantic Network relations as predicates. Processing relies on an underspecified syntactic analysis based on the SPECIALIST Lexicon (McCray *et al.*, 1994) and MedPost part-of-speech tagger (Smith *et al.*, 2004). MetaMap (Aronson, 2001) maps simple noun phrases to Metathesaurus concepts, and "indicator rules" map syntactic elements to Semantic Network predicates. For example, SemRep identifies the three semantic predications in (2) from the sentence fragment in (1):

(1) … dexamethasone is a potent inducer of multidrug resistance-associated protein expression in rat hepatocytes …

(2) Dexamethasone STIMULATES Multidrug Resistance Associated Proteins
Multidrug Resistance-Associated Proteins PART_OF Rats
Hepatocytes PART_OF Rats

These predications comprise executable knowledge and are amenable to further automatic manipulation.

3.2 Automatic Summarization

In the semantic abstraction paradigm of automatic summarization (Hahn and Mani, 2000) semantic predications serve as representation of the source text and are manipulated to generate a salient overview of input text. SemRep predications from multiple documents provide input to the Semantic MEDLINE summarizer, which provides a reduced and focused list of predications (a "semantic condensate").

Semantic condensates are based on a user-selected topic and a summarization perspective (Treatment of Disease, Substance Interactions, Diagnosis, or Pharmacogenomics). Each perspective is represented as a set of formal constraints on the arguments and the predicate of the input predications.

In all perspectives, the transformation from the initial list of predications to the reduced list in the semantic condensate is guided by four principles, which are informally defined as:

- *Relevance*: Include predications on the topic of the summary that conform to the selected summarization perspective
- *Connectivity*: Include additional useful predications on the basis of having shared arguments with the "relevant" predications
- *Novelty*: Eliminate, using UMLS hierarchical information, the predications the user already knows, identified as those having generic arguments, such as "Pharmaceutical Preparations" or "Disease"
- *Saliency*: Eliminate predications with low frequency of occurrence

4 System Implementation

4.1 Enhancing SemRep

SemRep had originally been developed with an emphasis on clinical research; it was enhanced for Semantic MEDLINE to accommodate linking the research literature to structured resources, including genetic databases. SemRep now augments mappings provided by MetaMap with ABGene (Tanabe *et al.*, 2002) and pattern matching to recognize and normalize gene names to Entrez Gene (Maglott *et al*., 2007). For example, MetaMap is unable to map the token "c-Jun" to a Metathesaurus concept; however, ABGene identifies it as a gene, and the normalization routine maps it to the Entrez Gene official symbol "JUN" and records its gene identifier (3725). The normalization mechanism uses a pre-computed index based on Entrez Gene official symbols, names, and aliases stored in a Berkeley DB table. The normalization index is updated periodically and is currently limited to human genes.

4.2 Semantic MEDLINE

Semantic MEDLINE is implemented as a three-tier, Java EE-based Web application (Fig. 1), which allows separation of user interface, application logic, and data storage. We leverage ma-

ture open-source technologies to the extent possible. The application runs in a Tomcat servlet container on an Apache http server and has been developed using the Apache Struts Web application framework (http://struts.apache.org/). This encourages the use of the MVC (Model-View-Controller) paradigm to provide a clean separation of application model, navigational code, and page design code through the use of Java Servlet API.

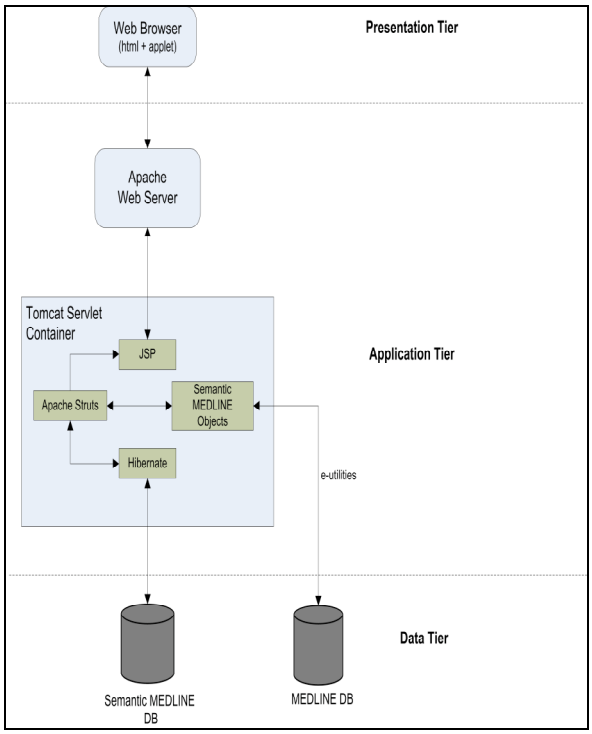

Fig. 1. Semantic MEDLINE architecture

A MySQL database is used to store Semantic MEDLINE data, which includes semantic predications extracted from MEDLINE citations in addition to UMLS Metathesaurus and Entrez Gene data. The database tables are pre-populated from plain text files that contain SemRep output and Metathesaurus/Entrez Gene data using Perl scripts. The Hibernate object/relational mapping (ORM) tool (http://www.hibernate.org/) provides enhanced database access through database connection pooling and query caching.

Semantic MEDLINE supports PubMed searching through NCBI's Entrez Programming Utilities API (http://eutils.ncbi.nlm.nih.gov/) to provide real-time access to PubMed records, retrieved and manipulated in XML format.

To visualize the semantic condensates as graphs in Semantic MEDLINE, we developed a Flash application using the Adobe Flex framework (http://www.adobe.com/products/flex) and the Flare visualization toolkit (http://flare.prefuse.org/), the ActionScript extension of the Prefuse toolkit written in Java. Nodes in a graph represent arguments in SemRep predications, and the arcs predicates. We enhanced the visualization capabilities provided by Flare by linking the semantic predications in the graph to external structured biomedical resources.

Arcs are linked to the MEDLINE citations from which the corresponding predications were extracted, while nodes are linked to three resources in addition to Entrez Gene: the UMLS Semantic Navigator (Bodenreider, 2000), Online Mendelian Inheritance in Man (OMIM) (Hamosh *et al.*, 2002), and Genetics Home Reference (GHR) (Mitchell *et al.*, 2004).

Linking to the UMLS Semantic Navigator uses Metathesaurus concept identifiers (CUI) and allows the user to view the context of a predication argument in the UMLS hierarchy. The gene name normalization procedure discussed above enables linking to Entrez Gene. OMIM identifiers are extracted from the OMIM *morbidmap* file periodically and associated with UMLS Metathesaurus concepts in the Semantic MEDLINE database, while GHR identifiers are extracted from GHR XML files periodically and, similarly, associated with Metathesaurus concepts.

SemRep is not fast enough to accommodate Semantic MEDLINE in real time. We therefore run SemRep on the MEDLINE database in an off-line process and store the extracted predications in the MySQL database as they become available. Currently, the database contains 9,224,765 predications extracted from 2,779,669 citations processed by MEDLINE during 2004, 2005, 2006, and 2007.

5 User Interface

The Semantic MEDLINE Web page has four tabs: Search, SemRep, Summarization and Visualization. The Search tab allows the user to specify a query and select PubMed limits. Titles of retrieved citations are displayed in tabular format, hyperlinked to PubMed. On this page and throughout, Semantic MEDLINE results can be saved in XML format for later reuse.

The SemRep tab presents predications extracted from citations retrieved. The user can then move to the Summarization tab and select a topic and perspective. Topics appear in a drop-down list, sorted by frequency of occurrence in the underlying SemRep predications.

Fig 2. A view from the Summarization tab

The user may also choose to disable filtering based on frequency of occurrence of predications (saliency filter). Fig. 2 shows a view from the Summarization tab.

The Visualization tab provides access to the graph representing the summarized semantic condensate, which guides navigation through the documents retrieved by the search. Nodes and arcs are color coded according to meaning. Node colors are determined by UMLS semantic groups (e.g. substances, procedures, disorders) (McCray *et al.*, 2001). The color legends for the nodes and arcs are displayed in the Filters tab on the right pane. Each item in the legends is a check box, and clicking on one of them shows or hides the nodes (or arcs) with that semantic type (or predicate) in the graph, providing focused views.

Clicking on a graph element displays information in the Information tab on the right pane. In addition to frequency of occurrence of the corresponding argument or predication, information for nodes includes UMLS concept identifier and semantic type for the corresponding argument as well as links (if available) to external resources, including the UMLS Semantic Navigator, Entrez Gene, GHR, and OMIM; for arcs, arguments of the corresponding predication and predicate name are given. The Citation button enables viewing the MEDLINE citations from which the predication was extracted, including PubMed identifier, title and abstract. The citation sentence in which the predication is asserted is highlighted. (See Fig.3 for some aspects of the visualization)

6 Evaluation

We have so far not conducted a user-centered evaluation. Accuracy of the predications generated by SemRep is crucial to overall effectiveness of Semantic MEDLINE. A summary of prior evaluations of SemRep and the automatic summarizer (see Table 1) suggests that average precision is near 77%. The evaluations conducted have generally been post-hoc and considered precision only; one study also assessed recall.

In each study, evaluation was limited to particular predicates: hypernymic (ISA) relations (Rindflesch and Fiszman, 2003), gene-disease etiological relations, such as CAUSE and PREDISPOSE, (Rindflesch *et al.,* 2003b) and finally, those relations focusing on pharmacogenomics, such as DISRUPTS and INHIBITS (Ahlers *et al.,* 2007a).

Evaluation of the automatic summarizer involved assessing accuracy of the predications in semantic condensates produced from various summarization perspectives. Two focused on treatment of disease (Fiszman *et al.*, 2004a; Fiszman *et al.*, 2004b), one with MEDLINE citations, and the other with an online medical encyclopedia as source documents. Semantic condensates regarding drug information were also evaluated (Fiszman *et al.*, 2006). All evaluation results are presented in Table 1.

7 Investigating Relaxin

We describe a scenario exploiting the components of Semantic MEDLINE to elucidate relaxin, a peptide hormone originally connected

Evaluation type and reference	Number of predications	Precision	Recall
SemRep			
Hypernymic (Rindflesch and Fiszman, 2003)	830	83%	
Gene-disease (Rindflesch *et al.*, 2003b)	1124	76%	
Pharmacogenomics (Ahlers *et al.*, 2007a)	623	73%	55%
Automatic summarizer			
Treatment of disease (Fiszman *et al.*, 2004a)	306	66%	
Treatment of disease (Fiszman *et al.*, 2004b)	190	87%	
Drug information (Fiszman *et al.*, 2006)	189	78%	
Total	**3262**	**77%**	

Table 1. SemRep/automatic summarization evaluation results

with parturition and more recently found to have a wider range of physiological implications. On the Search page, the user issues the query "relaxin" to PubMed, with the default dates 01/01/2004 through 12/31/2007 reflecting the part of MEDLINE currently available for processing. From PubMed Limits, accessible under "More options," "Abstracts" is selected. This query retrieves 349 citations, which generate 2899 predications (on the SemRep page). On the Summarization page the user chooses "Substance Interactions" as Summary Type and "Relaxin" as Summary Topic. The Saliency Filter (keeping only the most frequent predications) yields 119 predication tokens.

Summarized predications are displayed on the Visualization page as a graph, which provides an informative overview of the characteristics of relaxin as extracted from the retrieved citations. The user can also follow links to retrieve more detailed information on selected aspects of the graph. Contributing resources are the citations (linked to graph arcs) that produced the predications as well as related citations computed by PubMed. Additional structured knowledge sources include the UMLS Metathesaurus GHR, OMIM, and Entrez Gene (linked to graph nodes).

The current graph consists of 21 predication types with four predicate types: ISA, CAUSES, AFFECTS, and INTERACTS_WITH. One predication is disconnected from the main graph ("Isoproterenol CAUSES myocardium; injury"); the other 20 are connected with "Relaxin" as the central concept.

Hierarchical structure in the Metathesaurus, accessible from graph nodes, provides general information about the entities that relaxin is involved with. For example, two of these are shown to be peptides:

- Angiotensin II → Angiotensins → peptide hormone
- Adenylate Cyclase → Intracellular Signalling Peptides and Proteins → Peptides

Perusal of predicate types in the graph elucidates the major characteristics of relaxin in a principled way. "Relaxin" is in the following relationships:

- ISA: Hormones, peptide hormone
- CAUSES: Premature Birth
- AFFECTS: Renal fibrosis, Contraction, Apoptosis, Hemodynamics
- INTERACTS_WITH: Angiotensin II, Collagen, Progesterone, Adenylate Cyclase, Interleukin-11, RXFP1, RXFP2

Concentrating first on the ISA predications (extracted from 52 citations) provides an overview of relaxin function. For example, the first citation accessible from the arc between "Relaxin" and "Hormones" indicates an important relaxin function "…reverses cardiac and renal fibrosis…" (PMID 15967869), while the fourth (PMID 17266534) is a review article describing other relaxin characteristics: "…denoted initially as a hormone of pregnancy…" and "…many other physiological roles have been identified for relaxin, including cardiovascular and neuropeptide functions and an ability to induce the matrix metalloproteinases…" Further exploration of the graph reveals additional aspects of relaxin's activities. For example, clicking on the arc (ISA) between "Relaxin" and "peptide hormones" reveals a cognitive function for relaxin. The title of the first citation (PMID 16262650) is "Relaxin

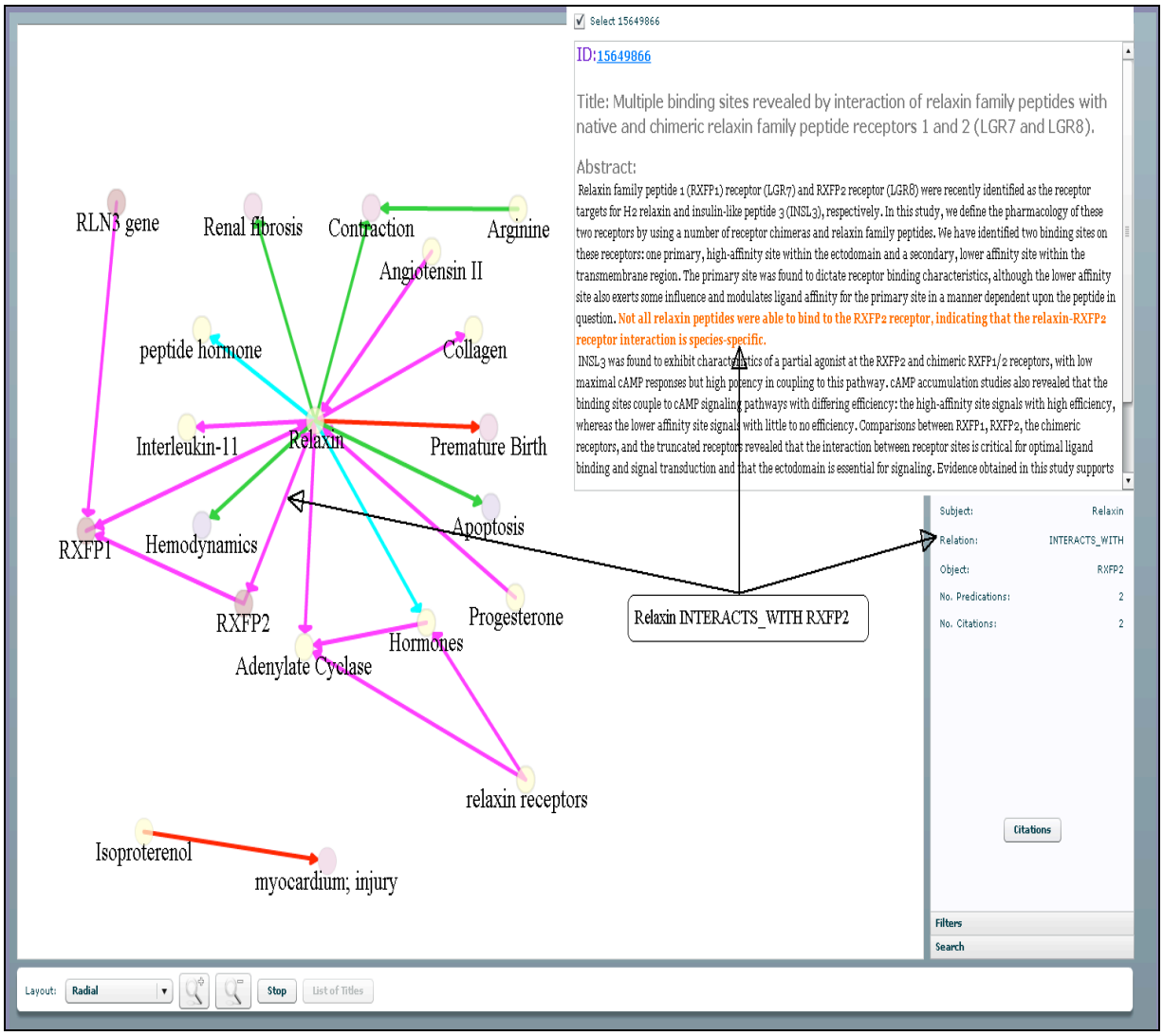

Fig. 3. Visualizing summarization results for Relaxin search, with Relaxin INTERACTS_WITH RXFP2 relation highlighted.

receptor activation in the basolateral amygdala impairs memory consolidation. ”

Information associated with the graph allows the user to pursue some particular aspect of relaxin in greater detail. For example, there are five citations available by clicking on the AFFECTS arc between “Relaxin” and “Hemodynamics.” Based on the known effects of relaxin during pregnancy, some of the basic research reported in these citations investigates its properties more generally. One of them (PMID 15198972), for example, tested “…whether relaxin can modify systemic arterial hemodynamics and load when chronically administered to nonpregnant rats,” while the goal of another (PMID 16172427) was “to determine the cardiovascular effects of rhRLX in hypertensive rats.” Another study (PMID 15271674) suggests practical implications: “…we speculate about the therapeutic potential of relaxin in renal and cardiovascular diseases.”

As noted above, SemRep precision is around 80%. A SemRep error in the graph is “relaxin receptors INTERACTS_WITH Hormones,” which was incorrectly extracted from two citations (PMID 14965317 and 15240635). Although neither asserts this predication, both publications may nonetheless be of interest regarding relaxin function. The title of the first is “Relaxin: new functions for an old peptide” and that of the second is “Increased expression of the relaxin receptor (LGR7) in human endometrium during the secretory phase of the menstrual cycle.”

The graph also serves as a guide to investigating the underlying mechanisms of relaxin. Interaction with two genes, RXFP1 and RXFP2, is shown. The title of one of the citations (PMID 15649866) that generated the predication assert-

ing interaction with RXFP2 confirms that these are the two major receptors for relaxin: "Multiple binding sites revealed by interaction of relaxin family peptides with native and chimeric relaxin family peptide receptors 1 and 2 (LGR7 and LGR8)."

Further exploration of RXFP2 is possible in Entrez Gene, which is accessible through a direct link from the RXFP2 node. Entrez Gene provides a wealth of technical information about this gene and its associated protein, including aliases (LGR8; GREAT; GPR106; INSL3R; LGR8.1; RXFPR2) and a brief summary. The functional information in the summary is augmented by Gene Ontology annotations and GeneRIFs (gene references into function), which are curated descriptive phrases culled from relevant MEDLINE citations. Finally, Entrez Gene provides links to a large number of structured knowledge sources, such as HGNC (HUGO Gene Nomenclature Committee) and KEGG (Kyoto Encyclopedia of Genes and Genomes).

Fig. 3 shows the visualization for the Relaxin search. One of the citations from which the predication "Relaxin INTERACTS_WITH RXFP2" is generated (PMID 15649866) is displayed, with the sentence that generates the predication highlighted.

8 Conclusion

We discussed the Semantic MEDLINE Web application, which helps PubMed users manage search results based on semantic natural language processing, automatic summarization, and visualization. To show its utility, we used the application as a guide in examining the peptide hormone relaxin, whose functions and mechanisms are not fully understood.

We are currently in the process of semantically analyzing the MEDLINE database and scaling the system without compromising performance. As the knowledge sources we rely on, including the UMLS and Entrez Gene, are continually updated, one challenge is to keep relevant data up-to-date. In addition, a large number of citations are added to MEDLINE daily, and these need to be made available through Semantic MEDLINE. At this time, we are putting in place procedures to automate data updating.

We are also exploring the extension of Semantic MEDLINE to supporting additional health-related textual databases, such as *ClinicalTrials.gov*. Finally, we plan to formally evaluate the user interface, which will no doubt lead to reassessing some of our design decisions and ultimate improvements in overall effectiveness of the application.

Acknowledgment

This research was supported in part by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

References

Ahlers, C., Fiszman, M., Demner-Fushman, D., Lang, F.-M., and Rindflesch. T. (2007a). Extracting semantic predications from Medline citations for pharmacogenomics. In *Proceedings of Pacific Symposium on Biocomputing,* 12:209-20.

Ahlers, C., Hristovski, D., Kilicoglu, H., and Rindflesch, T.C. (2007b) Using the Literature-Based Discovery Paradigm to Investigate Drug Mechanisms, In *Proceedings of AMIA Annual Symposium*, pp.6-10.

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

Blaschke, C., Andrade, M., Ouzounis, C., and Valencia, A. (1999) Automatic extraction of biological information from scientific text: protein-protein interactions, In *Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology*. pp. 60-7.

Bodenreider, O. (2000) A semantic navigation tool for the UMLS. In *Proceedings of AMIA Fall Symposium,* pp. 971.

Demner-Fushman, D., Lin, J. (2007) Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics,* 33 (1):63-103.

Doms, A. and Schroeder, M. (2005) GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Research*, 33(Web Server issue):W783-6.

Feldman, R., Regev, Y., Hurvitz, E., and Finkelstein-Landau, M. (2003) Mining the biomedical literature using semantic analysis and natural language processing techniques. *Biosilico*, 1(2):69-80.

Fiszman, M., Rindflesch, T.C., and Kilicoglu, H. (2004a) Abstraction summarization for managing the biomedical research literature. In *Proceedings of HLT-NAACL Workshop on Computational Lexical Semantics*. pp. 76-83.

Fiszman, M., Rindflesch, T.C., and Kilicoglu, H. (2004b) Summarization of an online medical encyclopedia. *MEDINFO*, 506-10.

Fiszman, M., Rindflesch, T.C., and Kilicoglu, H. (2006) Summarizing drug information in MEDLINE citations. In *Proceedings of AMIA Annual Symposium,* pp. 254-8.

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(Supplement 1):S74-82.

Fuller, S.S., Revere, D., Bugni, P., and Martin, G.M. (2004) A knowledgebase system to enhance scientific discovery: Telemakus. *Biomedical Digital Libraries*, 1(1):2.

Hamosh, A., Scott, A.F., Amberger, J., Bocchini, C., Valle, D., and McKusick, V.A. (2002) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 30(1):52-5.

Hahn, U. and Mani, I. (2000) The challenges of automatic summarization. *Compute*r, 33(11):29-36.

Hristovski, D., Friedman, C., Rindflesch, T.C., and Peterlin, B. (2006) Exploiting semantic relations for literature-based discovery. In *Proceedings of AMIA Annual Symposium,* pp. 347-53.

Jacquemart, P. and Zweigenbaum, P. (2003) Towards a medical question-answering system: a feasibility study. *Studies in Health Technology and Informatics*, 95:463-8.

Jensen, T.K., Laegreid, A., Komorowski, J., and Hovig, E. (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics*, 28(1):21-8.

Leroy, G., Chen, H. and Martinez, J.D. (2003) A shallow parser based on closed-class words to capture relations in biomedical text. *Journal of Biomedical Informatics*, 36(3):145-58.

Lindberg, D. A., Humphreys, B. L., and McCray, A. T. (1993) The Unified Medical Language System. *Methods of Information in Medicine.* 32(4):281–91.

Lussier, Y., Borlawsky, T., Rappaport, D., Liu, Y., and Friedman C. (2006) PhenoGO: assigning phenotypic context to Gene Ontology annotations with natural language processing. In *Proceedings of Pacific Symposium on Biocomputing*, pp. 64-75.

Maglott, D., Ostell, J., Pruitt, K.D., and Tatusova, T. (2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*, 35(Suppl):D26-31.

McCray, A.T., Burgun, A., and Bodenreider, O. (2001). Aggregating UMLS semantic types for reducing conceptual complexity. *MEDINFO*, 10(Pt 1):216-20.

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

McKeown, K.R., Chang, S.-F., Cimino, J., Feiner, S., Friedman, C., Gravano, L., Hatzivassiloglou, V., Johnson, S., Jordan, D.A., Klavans, J. L., Kushniruk, A., Patel, V., and Teufel, S. (2001) PERSIVAL, a system for personalized search and summarization over multimedia healthcare information. In *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 331-40.

Mitchell, J. A., Fun, J., and McCray, A.T. (2004) Design of Genetics Home Reference: A new NLM consumer health resource. *Journal of the American Medical Informatics Association.* 11(6):439-47.

Plake, C., Schiemann, T., Pankalla, M., Hakenberg, J., and Leser, U. (2006) ALIBABA: PubMed as a graph. *Bioinformatics*, 22(19): 2444-5.

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 AMIA Annual Symposium,* pp. 554-8.

Sable, C., Lee, M., Zhu, H.R., and Yu, H. (2005) Question analysis for biomedical question answering. In *Proceedings of AMIA Annual Symposium*, pp. 1102.

Smith, L., Rindflesch, T.C., and Wilbur, W.J. (2004) MedPost: a part-of-speech tagger for biomedical text. *Bioinformatics,* 20(14):2320-1.

Sneiderman, C.A., Demner-Fushman, D., Fiszman, M., Ide, N., and Rindflesch, T.C. (2007) Knowledge-based methods for helping clinicians find answers in MEDLINE. *Journal of the American Medical Informatics Association*, 14(6):772-80.

Srinivasan, P. and Libbus, B. (2004) Mining MEDLINE for implicit links between dietary substances and diseases. *Bioinformatics*, 20(Suppl 1):I290–I296.

Swanson, D.R. (1986) Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, 30(1):7-18.

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

Tao, Y., Friedman, C., and Lussier, Y.A. (2005) Visualizing information across multidimensional post-genomic structured and textual databases. *Bioinformatics*, 21(8):1659-67.

Wedgwood, J. (2005) MQAF: a medical question-answering framework. In *Proceedings of AMIA Annual Symposium*, pp. 1150.