Information Services & Use 31 (2011) 15–21
DOI 10.3233/ISU-2011-0627
IOS Press

Semantic MEDLINE: An advanced information management application for biomedicine

Thomas C. Rindflesch *, Halil Kilicoglu, Marcelo Fiszman, Graciela Rosemblat and Dongwook Shin

Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

Abstract. To support more effective biomedical information management, Semantic MEDLINE integrates document retrieval, advanced natural language processing, automatic summarization and visualization into a single Web portal. The application is intended to help manage the results of PubMed searches by condensing core semantic content in the citations retrieved. Output is presented as a connected graph of semantic relations, with links to the original MEDLINE citations. The ability to connect salient information across documents helps users keep up with the research literature and discover connections which might otherwise go unnoticed. Semantic MEDLINE can make an impact on biomedicine by supporting scientific discovery and the timely translation of insights from basic research into advances in clinical practice and patient care. Semantic MEDLINE is illustrated here with recent research on the clock genes.

Keywords: Biomedical information management, semantic processing, automatic summarization, graphical information representation, clock genes

1. Introduction

Access to online text is provided by document retrieval systems such as Google and PubMed (for biomedical information). The underlying technology of these systems typically manipulates text strings, perhaps augmented by frequency of occurrence and distribution patterns, and has remained largely unchanged since the 1980s [21]. Such systems have no access to the meaning of the text being processed. In biomedical information management, for example, more effective language processing is needed to support emerging applications, such as text mining aimed at task-driven extraction of facts to observe trends [2], those that connect text and structured data [8], question answering systems [4] and literature based discovery [24], which provides assistance to scientific research.

Automatic semantic interpretation (e.g., [7]) is intended to augment document retrieval systems by manipulating information, not just documents, and thereby bridge the gap between text and meaning. In the biomedical domain the Semantic MEDLINE application [14] exploits this technology to provide enhanced access to the research literature. The application calls on PubMed to return MEDLINE citations

* Corresponding author: Thomas C. Rindflesch, PhD, Lister Hill National Center for Biomedical Communications, National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894, USA. Tel.: +1 301 435 3191; Fax: +1 301 496 0673; E-mail: tcr@nlm.nih.gov.

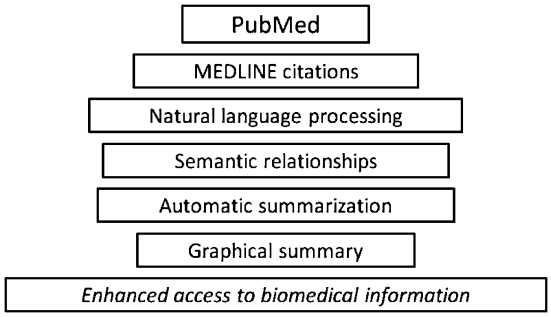

Fig. 1. Semantic MEDLINE processing.

in response to a user's query. Semantic processing extracts relationships representing the meaning of retrieved citations. After automatic summarization to focus on salient information, results are presented to the user as a connected graph of relationships for further exploration (Fig. 1).

2. Semantic interpretation of biomedical text: SemRep

Semantic MEDLINE calls on the SemRep natural language processing system [18,19] to extract semantic relationships. SemRep inspects each sentence in input text and identifies some of the relationships representing the meaning of the sentence. Semantic relationships are represented as predications, a formal representation having a predicate and arguments. For example, in Fig. 2, the predication "Genes AFFECTS Circadian Rhythms" was extracted from the sentence "Clock genes are the genes that control circadian rhythms in physiology and behavior". SemRep semantic predications provide a normalized representation of (part of) the meaning of a sentence and can be further manipulated by computational means.

SemRep depends on domain knowledge in the Unified Medical Language system (UMLS) developed by the US National Library of Medicine [12]. The UMLS consists of three knowledge sources: the SPECIALIST Lexicon, Metathesaurus and Semantic Network [6]. SemRep exploits the latter two for domain knowledge. The Metathesaurus contains a very large number of concepts compiled from an array of knowledge sources in biomedicine, focused on the clinical domain. A Metathesaurus concept consists of a collection of synonyms drawn from constituent vocabularies and terminologies. One of these is selected as the concept name. Concepts cover areas such as diseases, physiologic processes, drugs, anatomical concepts, organisms and population groups and concept classes that represent the same areas are grouped into semantic types, such as "Disease or Syndrome", "Pharmacologic Substance" and "Population Group".

A SemRep predication has UMLS Metathesaurus concepts as arguments and a UMLS Semantic Network relation as predicate. In the example given earlier, both arguments ("Genes" and "Circadian Rhythms") are Metathesaurus concepts, and the predicate ("AFFECTS") is a relation in the Semantic Network that binds the two arguments in this instance. During processing, the underlined text is mapped to the arguments and the italicized verb is mapped to the predicate. The MetaMap program [3], is used to map text to Metathesaurus concepts. Each concept has a semantic type associated with it, as shown in the following example:

Concept name: Genes; Semantic type: Gene or Genome.
Concept name: Circadian Rhythms; Semantic type: Organism Function.

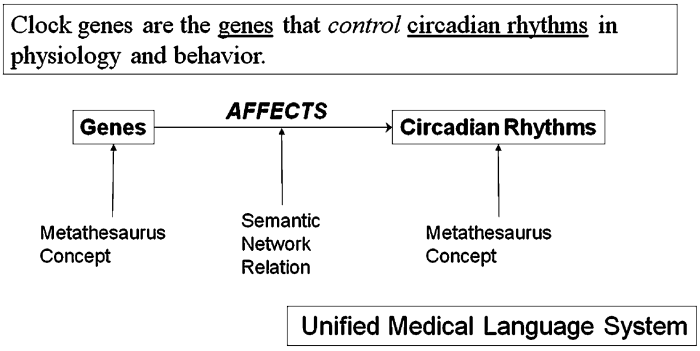

Fig. 2. Example predication extracted from text by SemRep.

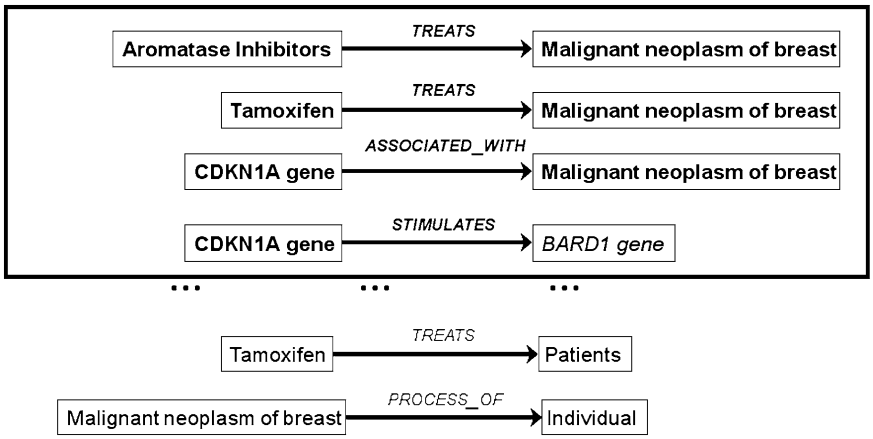

Fig. 3. SemRep predications extracted from text on breast cancer. Predications outside the box are eliminated by automatic summarization.

The UMLS Semantic Network stipulates relationships between concepts, stated in terms of the semantic types assigned to Metathesaurus concepts. These relationships form a pattern for the semantic predications that SemRep is allowed to identify and extract from biomedical text. That is, the semantic types of the Metathesaurus concepts serving as argument of SemRep predications must match those for the specific relationships in the Semantic Network, such as one of the following:

Gene or Genome AFFECTS Organism Function.
Therapeutic or Preventive Procedure USES Medical Device.
Pharmacologic Substance TREATS Disease or Syndrome.
Body Location or Region LOCATION_OF Biologic Function.
Disease or Syndrome OCCURS_IN Population Group.
Disease or Syndrome PROCESS_OF Organism.

SemRep extracts an array of predicate types in the biomedical domain. For example, the predications in Fig. 3 were extracted from text on the topic of breast cancer, and provide information on various aspects of this disease. The predications inside the box above are more informative that those below. Automatic summarization [9] is used to eliminate the less informative predications. A major aspect of

this processing is to use the UMLS to identify concepts that are too general to be informative, such as "Patient" and "Individual". Predications having such concepts as arguments are eliminated.

Regarding domain coverage, SemRep was initially developed for clinical medicine (the major focus of the UMLS), and has been extended to genetic etiology and substance interactions [20], pharmacogenomics and molecular biology [1], as well as influenza epidemic preparedness [13]. Research currently addresses extending the program to documents on public health and climate and health. Consideration is being given to extending coverage beyond biomedicine, to biomedical informatics, which includes practical application in the computer science domain.

SemRep has been run on MEDLINE, a bibliographic database of the biomedical research literature compiled and maintained by the US National Library of Medicine, containing some 19 million citations (1940s to present). 7.6 million citations (dated 01/01/99 through 02/28/11) have been processed and more than 26 million semantic predications extracted. These are stored in an SQL database and RDF triple store, and made available to the research community. SemRep semantic predications support the Semantic MEDLINE application.

3. Semantic MEDLINE application

We illustrate the use of Semantic MEDLINE by searching on the clock genes, discovered in 1971 in fruit flies [23] and subsequently found in humans (and all organisms). Originally these genes were thought to control sleep-related circadian rhythms only [16]. However, recent research provides insight into their physiological consequences underpinning the etiology of a range of common disorders.

In creating and exploiting a graphical summary in Semantic MEDLINE, the first step is to issue a PubMed query, in this case: "clock gene". We limit results to the 1000 most recent citations, and apply SemRep, which produces 404 semantic predications. These are summarized and displayed in a graph of nodes and arcs representing arguments and predicates (Fig. 4). The nodes are color coded for UMLS semantic type. For example in Fig. 4, diseases, such as "Metabolic Diseases" (about 10 o'clock in the central portion of the graph) are pink; physiologic functions (e.g., "Homeostasis" at 11 o'clock) are gray; and genes ("AANAT" at 1 o'clock) are mauve. Arcs representing predicates are also color coded. (Predications are read in the direction of the arrow.) For example cyan is used for the relation predisposes: "CLOCK PREDISPOSES Metabolic Diseases" (11 o'clock); magenta for INTERACTS_WITH: "CLOCK INTERACTS_WITH Phosphotransferases" (11 o'clock); and green for "AFFECTS": "CLOCK AFFECTS Homeostasis" (12 o'clock).

The graph summarizes and organizes the content of the documents processed and can be exploited in two ways: by taking advantage of both its informative and indicative aspects. In summarization theory, "informative" refers to aspects of the summary itself, while "indicative" refers to the way in which additional information is provided.

Inspection of the nodes in Fig. 4 reveals concepts central to research on the clock genes. For example, the major genes (CRY1, CRY2, PER1, PER2, DEC1, ARNTL, ARNTL2 and CLOCK gene) appear in the graph along with several disorders affected by them (tumor growth, Mood Disorders, Depressive symptoms, Metabolic syndrome, Winter Depression). Underlying physiologic functions (Physiological aspects, Circadian Rhythms, Growth) are seen as well. The relationships in which these concepts participate provide more specific information. For example, it is seen that CLOCK gene is ASSOCIATED_WITH (brown) Depressive symptoms (lower right section of the figure) and DISRUPTS (yellow) tumor growth (right-hand section of the figure).

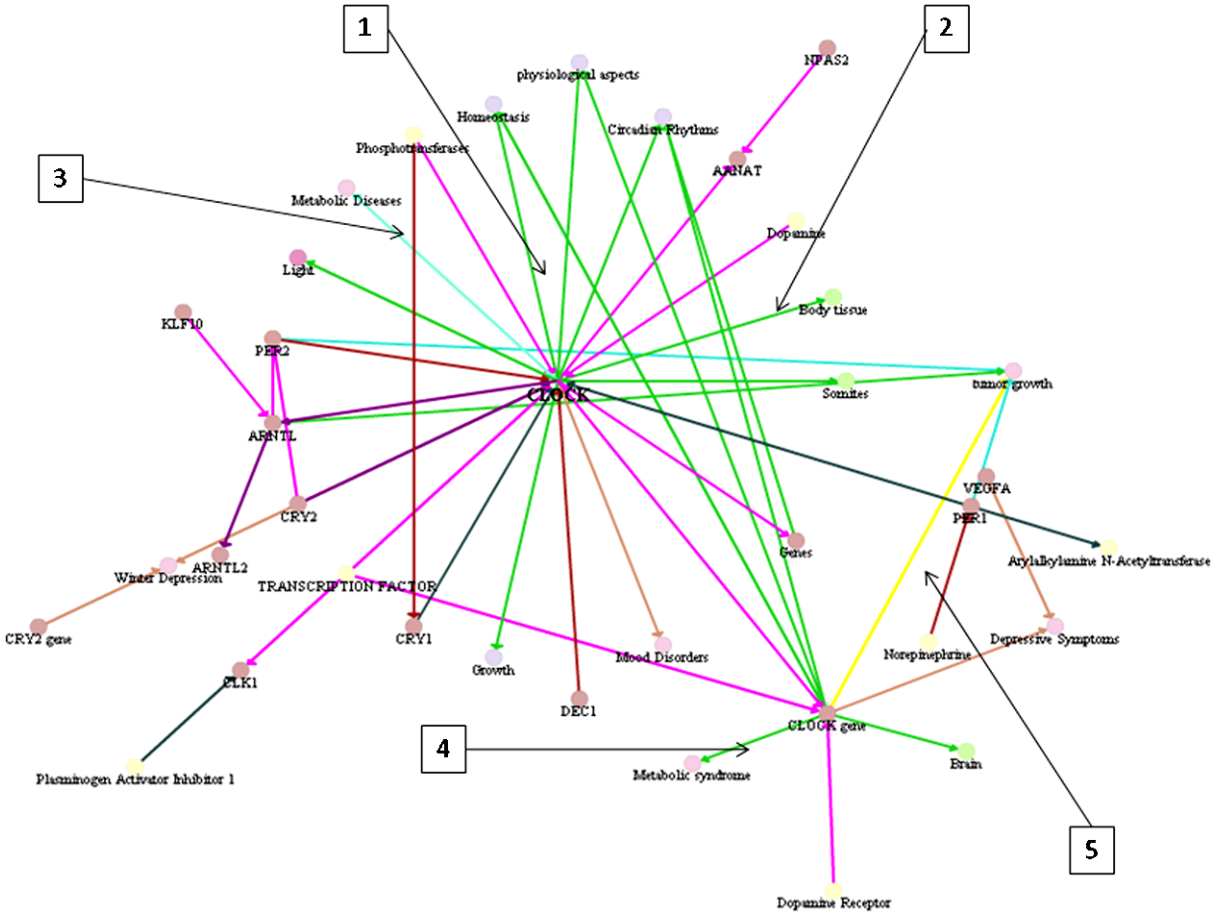

Fig. 4. Semantic MEDLINE graph showing SemRep predications extracted from MEDLINE citations retrieved with the PubMed query "clock gene". (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/ISU-2011-0627.)

More information is available by clicking on the arcs, which are linked to the MEDLINE citations from which the relationships were extracted. This facility allows the user to investigate relevant phenomena in more depth. For example, from selected relationships, three major aspects of the clock genes (as reflected in current research) can be determined:

(1) The clock genes affect core aspects of physiology, especially metabolism.
(2) The clock genes can be influenced by various factors, including body substances (e.g., hormones, cytokines), food and cognitive state.
(3) Disruption of the clock genes can contribute to the development of metabolic disease.

The relevant relationships are labeled in Fig. 4. Text in each of the abstracts from which the relationship was extracted illustrates the corresponding points noted.

- *CLOCK* AFFECTS *homeostasis* (labeled "1" in Fig. 4) "The circadian clock controls energy homeostasis by regulating circadian expression of proteins involved in metabolism" [5].
- *CLOCK* AFFECTS *Body tissue* (labeled "2") "This review will focus on the interconnection between the circadian clock and metabolism, with implications for obesity and how the circadian clock is

influenced by hormones, nutrients, and timed meals" [10].

- *CLOCK* PREDISPOSES *Metabolic Diseases* (labeled "3") "Emerging evidence suggests that circadian clock function is closely linked to metabolic homeostasis and that rhythm disruption can contribute to the development of metabolic disease" [15].

The Semantic MEDLINE graph supports the kind of patterns that underpin literature-based discovery [22]. One such pattern characterizes closed discovery, in which a relationship A → C can be explicated by finding two relationships, A → B and B → C, thereby providing B as a mechanistic explanation. As an example, we consider cancer and obesity. There has been a longstanding and pervasive research interest in the association between these two phenomena. The PubMed query "obesity AND (cancer OR neoplasm)" returns 11,091 citations. The earliest is from 1947 [17]. However, until recently, potential mechanisms involved have not been well understood. The clock genes constitute one such mechanism, which is partly explicated in Fig. 4. We first look at the citation [11] from which the predication "CLOCK gene AFFECTS Metabolic syndrome" (labeled "4" in Fig. 4) was extracted. The thrust of the research reported is that ". . .obesity induced by high-fat diet alters the circadian-clock system. . ." and further that "expressions of circadian-clock genes and circadian clock-controlled genes, including Per1–3. . . were altered in the livers and/or kidneys".

With that in mind, we next inspect the predication "CLOCK gene DISRUPTS tumor growth" (labeled "5") and the citation [25] from which it was generated. The citation contains two statements that implicate Per1 and Per2 in cancer: "Per2 is a core clock gene, the product of which suppresses cancer cell proliferation and tumor growth *in vivo* and *in vitro*". And "Down-regulation of the expression of tumor Per1 increases cancer cell growth *in vitro* and tumor growth *in vivo*. . .". Taken together, these two citations provide the clock genes as a mechanistic link (B) for the observed relationship between cancer and obesity (A → C). Although this is not an actual discovery, it illustrates the kind of methodology Semantic MEDLINE supports as a resource for biomedical researchers.

4. Conclusion

Emerging applications in biomedical information management require more expressive text analysis than that provided by available document retrieval systems. Semantic MEDLINE provides enhanced access to the biomedical research literature by combining PubMed document retrieval, semantic relationships, and automatic summarization. Results are presented to the user in a graph representing an overview of content with links to source documents. The core technology for this application is the SemRep natural language processing system, which extracts semantic information from biomedical text, supported by domain knowledge in the Unified Medical Language system. Semantic MEDLINE usage is illustrated with recent research literature on the clock genes, showing the value of this system for keeping abreast of research trends and suggesting the possibility of exploiting it for scientific discovery.

Acknowledgement

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

References

[1] C.B. Ahlers, M. Fiszman, D. Demner-Fushman, F. Lang and T.C. Rindflesch, Extracting semantic predications from MEDLINE citations for pharmacogenomics, in: *Pacific Symposium on Biocomputing*, World Scientific, Singapore, 2007, pp. 209–220.

[2] S. Ananiadou and J. McNaught, eds, *Text Mining for Biology and Biomedicine*, Artech House Books, London, 2006.

[3] A.R. Aronson and F.M. Lang, An overview of MetaMap: historical perspective and recent advances, *Journal of the American Medical Informatics Association* **17** (2010), 229–236.

[4] S.J. Athenikos and H. Han, Biomedical question answering: a survey, *Computer Methods and Programs in Biomedicine* **99** (2010), 1–24.

[5] M. Barnea, Z. Madar and O. Froy, High-fat diet followed by fasting disrupts circadian expression of adiponectin signaling pathway in muscle and adipose tissue, *Obesity (Silver Spring)* **18** (2010), 230–238.

[6] O. Bodenreider, The unified medical language system (UMLS): integrating biomedical terminology, *Nucleic Acids Research* **32** (2004), D267–D270.

[7] J. Bos and S. Pulman, eds, *Proc. 9th International Conference on Computational Semantics (IWCS)* and *International Workshop on Computational Semantics*, Oxford, 2011.

[8] K.B. Cohen, O. Bodenreider and L. Hirschman, Linking biomedical information through text mining: session introduction, in: *Pacific Symposium on Biocomputing*, World Scientific, Singapore, 2006, pp. 1–3.

[9] M. Fiszman, T.C. Rindflesch and H. Kilicoglu, Abstraction summarization for managing the biomedical research literature, in: *Proc. HLT-NAACL Workshop on Computational Lexical Semantics*, Association for Computational Linguistics, Stroudsburg, 2004, pp. 76–83.

[10] O. Froy, Metabolism and circadian rhythms – implications for obesity, *Endocrine Reviews* **31** (2010), 1–24.

[11] M.C. Hsieh, S.C. Yang, H.L. Tseng, L.L. Hwang, C.T. Chen and K.R. Shieh, Abnormal expressions of circadian-clock and circadian clock-controlled genes in the livers and kidneys of long-term, high-fat-diet-treated mice, *International Journal of Obesity* **34** (2010), 227–239.

[12] B.L. Humphreys, D.A. Lindberg, H.M. Schoolman and G.O. Barnett, The unified medical language system: an informatics research collaboration, *Journal of the American Medical Informatics Association* **5** (1998), 1–11.

[13] A. Keselman, G. Rosemblat, H. Kilicoglu, M. Fiszman, H. Jin, D. Shin and T.C. Rindflesch, Adapting semantic natural language processing technology to address information overload in influenza epidemic management, *Journal of the American Society for Information Science and Technology* **61** (2010), 2531–2543.

[14] H. Kilicoglu, M. Fiszman, A. Rodriguez, D. Shin, A.M. Ripple and T.C. Rindflesch, Semantic MEDLINE: a web application to manage the results of PubMed searches, in: *Proc. 3rd International Symposium in Semantic Mining in Biomedicine*, European Bioinformatics Institute, Hinxton, 2008, pp. 69–76.

[15] J. Kovac, J. Husse and H. Oster, A time to fast, a time to feast: the crosstalk between metabolism and the circadian clock, *Molecules and Cells* **28** (2009), 75–80.

[16] P.L. Lowrey and J.S. Takahashi, Mammalian circadian biology: elucidating genome-wide levels of temporal organization, *Annual Review of Genomics and Human Genetics* **5** (2004), 407–441.

[17] W.T. Moss, Common peculiarities of patients with adenocarcinoma of the endometrium with special reference to obesity, body build, diabetes and hypertension, *American Journal of Roentgenology and Radium Therapy* **58** (1947), 203–210.

[18] T.C. Rindflesch and M. Fiszman, The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text, *Journal of Biomedical Informatics* **36** (2003), 462–477.

[19] T.C. Rindflesch, M. Fiszman and B. Libbus, Semantic interpretation for the biomedical research literature, in: *Medical Informatics: Knowledge Management and Data Mining in Biomedicine*, H. Chen, S.S. Fuller, C. Friedman and W. Hersh, eds, Springer, New York, 2005, pp. 399–422.

[20] T.C. Rindflesch, B. Libbus, D. Hristovski, A.R. Aronson and H. Kilicoglu, Semantic relations asserting the etiology of genetic diseases, in: *Proc. AMIA Annual Symposium*, American Medical Informatics Association, Bethesda, MD, 2003, pp. 554–558.

[21] D. Satterthwaite, Emerging technologies to speed information access, *Information Services & Use* **30** (2010), 99–105.

[22] D.R. Swanson, Fish oil, Raynaud's syndrome, and undiscovered public knowledge, *Perspectives in Biology and Medicine* **30** (1986), 7–18.

[23] J.S. Takahashi, Finding new clock components: past and future, *Journal of Biological Rhythms* **19** (2004), 339–347.

[24] M. Weeber, J.A. Kors and B. Mons, Online tools to support literature-based discovery in the life sciences, *Briefings in Bioinformatics* **6** (2005), 277–286.

[25] X. Yang, P.A. Wood, C.M. Ansell, D.F. Quiton, E.Y. Oh, J. Du-Quiton and W.J. Hrushesky, The circadian clock gene Per1 suppresses cancer cell proliferation and tumor growth at specific times of day, *Chronobiology International* **26** (2009), 1323–3139.