Marc Weeber, Jan A. Kors and Barend Mons
are all part of the Biosemantics group (http://www.biosemantics.org) at Erasmus MC, Rotterdam, The Netherlands. The group's main interest is to develop and apply novel computational biology technology to support discovery in the life sciences.

Online tools to support literature-based discovery in the life sciences

Marc Weeber, Jan A. Kors and Barend Mons

Date received (in revised form): 7th June 2005

Abstract

In biomedical research, the amount of experimental data and published scientific information is overwhelming and ever increasing, which may inhibit rather than stimulate scientific progress. Not only are text-mining and information extraction tools needed to render the biomedical literature accessible but the results of these tools can also assist researchers in the formulation and evaluation of novel hypotheses. This requires an additional set of technological approaches that are defined here as literature-based discovery (LBD) tools. Recently, several LBD tools have been developed for this purpose and a few well-motivated, specific and directly testable hypotheses have been published, some of which have even been validated experimentally. This paper presents an overview of recent LBD research and discusses methodology, results and online tools that are available to the scientific community.

Keywords: *literature-based discovery, text mining, Unified Medical Language System*

INTRODUCTION

Scientific discovery is a typical human intellectual activity. Based on observations and theory, researchers define hypotheses that they test experimentally. The experimental outcome may lead to modification or even falsification of the hypotheses. There has been a considerable interest in the computational support of human researchers in both defining and testing hypotheses.[1] Indeed, such support is needed as the amount of experimental data and scientific information is overwhelming and may soon inhibit rather than stimulate progress. Informatics tools may be fruitfully integrated in the practice of hypothesis-driven scientific research.[2] Only recently have fully automatic and integrated approaches of hypothesis generation and experimental testing started to appear.[3]

When defining new hypotheses, scientists combine observations and existing knowledge in a novel way. Thus, first of all, keeping abreast with existing and emerging knowledge is important. For biomedicine, this means that every researcher has to be proficient in using MedLine, the main repository of published biomedical literature. Still, a standard query to MedLine retrieves only explicit knowledge that pertains to the query. Implicit knowledge that can only be inferred from existing knowledge is not available through standard tools, but may be very valuable when conducting research.

Consider, for example, myasthenia gravis, which is an organ-specific autoimmune disease aimed at the nicotine acetylcholine receptor at the neuromuscular junctions. The pathophysiology of myasthenia gravis is complex and at present mostly unknown.[4] To find treatments for this disease that are novel, it is principally impossible to query MedLine, or any other database for that matter, for an answer. However, it is possible to try to find implicit knowledge that may guide the process of finding promising treatments. It is known, for instance, that aberrant production of, among others, TNF-alpha, IL-10 and IL-12 are considered to play a role in the pathology.[5,6] On the other hand, thalidomide (softenon), a former sedative and currently an immunomodulatory

Marc Weeber,
Department of Medical Informatics,
Erasmus MC – University Medical
Center Rotterdam,
PO Box 1738,
3000 DR Rotterdam, The
Netherlands

Tel: +31 10 4088118
Fax: +31 10 4089447
E-mail: m.weeber@erasmusmc.nl

agent, has strong inhibitory effects on mononuclear cell production of IL-12,[7] has a stimulatory effect on IL-10 production,[8] and degrades TNF-alpha mRNA.[9] By putting these different pieces of knowledge together, the novel hypothesis emerges that thalidomide may treat myasthenia gravis.[10]

Hypothesis generation

The process of (semi-)automatically inferring implicit knowledge from literature databases, which results in well-motivated and testable hypotheses, is called literature-based discovery (LBD). In this paper, a short background on LBD and an overview of recent discoveries are first provided. Subsequently, the different techniques employed are discussed and the freely available online tools that have been developed to assist LBD are reviewed.

LITERATURE-BASED DISCOVERY

In 1986, Don Swanson presented his first literature-based hypothesis that fish oil may have beneficial effects in patients with Raynaud's disease.[11] Fish oil lowers blood viscosity, inhibits platelet aggregation and causes vascular reactivity. On the other hand, patients with Raynaud's disease have increased blood viscosity and platelet aggregation and suffer from impaired vascular reactivity. In 1986, no one had made this implicit link explicit until Swanson connected the apparently disconnected fields of biomedical expertise. The possibility of linking different scientific disciplines through intermediate, or shared, interests has commonly been described as Swanson's *ABC* model (Figure 1).

Swanson's ABC model

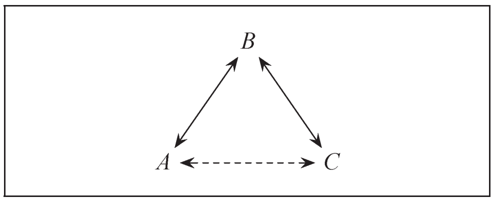

Figure 1: Swanson's *ABC* model of discovery. If *A* and *B* are related, and *B* and *C* are related, it follows that *A* and *C* might be indirectly related

Swanson's *ABC* model can be implemented as two different discovery processes, see Weeber *et al.*[12] for an overview. An open discovery process is characterised by the generation of a hypothesis, a closed discovery process by the elaboration of a hypothesis.

Figure 2 depicts the open approach starting with disease *C*. Interesting clues (*B*) about the mechanism of the disease will be sought in the literature. In terms of Swanson's first discovery, the problem is to find underlying (patho)physiological mechanisms of Raynaud's disease. For the most interesting clues, substances (*A*) are looked for that may interact with these mechanisms. Swanson focused on dietary factors that may have an influence on relevant *B* processes. In the discovery process, it is likely that many *B*s and *A*s will be found. In fact, the challenge of discovery support tools is to constrain the vast amount of possibilities. As the result of an open discovery process, one may formulate the specific hypothesis that substance *A* can be used for the treatment of disease *C* via one or more *B* pathways.

Figure 3 depicts the closed discovery approach of verifying and elaborating an initial hypothesis, for instance, the treatment of disease *C* with substance *A*.

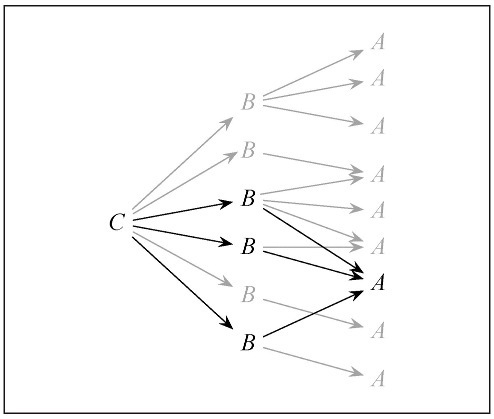

Figure 2: Open discovery process or generating a hypothesis. The search starts at *C*, for instance a disease, and results in *A*, possibly a drug. The intermediate *B* steps may represent (patho)physiological mechanisms. The black arrows indicate potentially interesting pathways of discovery, the grey ones pathways that do not qualify

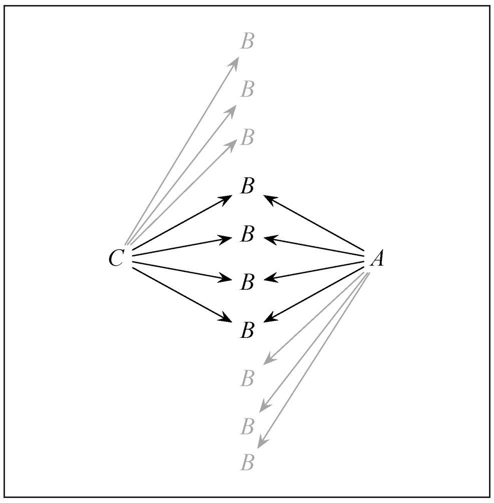

Figure 3: Closed discovery process or evaluating a hypothesis. The search process starts simultaneously from *C* (eg disease) and *A* (drug), resulting in overlapping *B*s (potential mechanisms). The black arrows indicate potentially interesting pathways of discovery, the grey ones spurious links. The more pathways are found, the stronger the support for the hypothesis

Information on common mechanistic processes (*B*) are extracted from the literature. Typically, the more pathways between *A* and *C* are extracted, the stronger the hypothesis will be. An example relates to the observation that in patients with multiple myeloma who were treated with thalidomide, two responders had a concomitant improvement in chronic hepatitis C.[13] A closed discovery process may elucidate possible underlying mechanisms of how thalidomide may treat chronic hepatitis C.[10]

Literature-based discoveries are concise, well motivated and testable: a specific *C* is connected to a specific *A*. In reaching this specificity, intermediate steps are often not that precise. For instance, when looking for genes (typical *A*s), involved in a specific disease *C*, the actual approach is a hybrid of a closed and open discovery. This aspect may be used to conduct an open-like search with tools that only support closed discoveries.

Text mining and discovery

Literature-based discovery should be regarded as a technique employed subsequent to text mining for explicit facts. General text-mining approaches focus on the extraction of relevant entities, eg genes and proteins, and relationships between them, eg protein–protein interactions. The users of these approaches generally retrieve known, explicit co-occurrence-based knowledge that they personally were not always aware of and text mining can be viewed as an efficient way of keeping abreast with the most important facts in the literature. Building largely on mined entities and facts, LBD tools attempt to combine extracted information into serendipitous and truly novel hypotheses. Because many combinations of mined facts are possible, the main aim of LBD systems is to confine the explosively growing number of possible hypotheses to those that have the highest probability to be consistent. Most systems provide a list of hypotheses rank-ordered according to certain likelihood, for instance, the number of intermediate *B* concepts of a certain *AC* hypothesis. In contrast to text-mining and information extraction approaches, there is no straightforward evaluation possible as it is not easy to establish the correctness of generated hypotheses.

RECENT LITERATURE-BASED DISCOVERIES

Novel potential therapeutic applications for thalidomide

Since 1986, Swanson and his co-worker Smalheiser have suggested several different hypotheses in the biomedical domain, and some of them have been corroborated experimentally. Smalheiser and Swanson[14] and Srinivasan[15] provide overviews. The recent years have seen an increased interest in literature-based discovery and, indeed, several new literature-based hypotheses have been published in biomedical journals. Weeber *et al.* have proposed four new therapeutic applications for the drug thalidomide: myasthenia gravis, chronic hepatitic C, *Helicobacter pylori*-induced gastritis and acute pancreatitis.[10] Similarly, Srinivasan and Libbus[16] have suggested novel therapeutic uses for the dietary substance curcuma longa: retinal diseases, Crohn's disease and disorders related to the spinal cord. Wren and colleagues[17] have suggested that chlorpromazine may reduce cardiac hypertrophy. They have also observed that the pathogenesis of non-insulin dependent diabetes is most likely epigenetic.[18] Hristovski *et al.*[19] have found novel candidate genes that may be involved in bilateral perisylvian polymicrogyria.

Experimental support for literature-based discoveries

Most literature-based proposed hypotheses are well motivated, very specific and directly testable. Indeed, Wren *et al.* showed in a rodent model that chlorpromazine reduced isoprotenerol-induced cardiac hypertrophy.[17] Animal models of the studied diseases are often available, and a first proof of principle is feasible. Also, when candidate disease genes are identified, mutation screening of these genes in patients should be straightforward.

EMPLOYED TEXT-MINING TECHNIQUES

While there have been a few literature-based discoveries outside biomedicine,[20,21] all literature-based discovery approaches discussed in this paper have a biomedical focus and use textual information derived from MedLine. Most often used are titles, abstracts and MeSH headings. Hristovski uses additional, non-textual information in the form of chromosome location and gene expression localisation extracted from LocusLink,[19,22] similarly to research by Perez-Iratxeta *et al.*[23]

Before algorithms to discover implicit knowledge can be meaningfully applied, the explicit facts from literature have to be mined as efficiently as possible. This preprocessing step is crucial and many studies have been conducted in this field. However the details of these studies fall outside the scope of this review. For an overview of the state of the art, we refer to the recent BMC special issue on BioCreAtIvE, the Critical Assessment of Information Extraction in Biology.[24] Here, only the methods employed by the LBD tools under review are briefly mentioned.

Swanson's original discoveries were based on an exhaustive reading of the literature; however, Swanson started developing text analysis scripts that evolved into the Arrowsmith system.[25] Arrowsmith's basic approach is to identify the longest overlapping terms between titles from literature sets containing A and C query terms. The overlap between the titles *Raynaud's phenomenon and blood viscosity* and *Beneficial effect of fish oil on blood viscosity in peripheral vascular disease*, for instance, is the term *blood viscosity*. Gordon and Lindsay, the first to follow Swanson, developed a more sophisticated methodology based on lexical statistics.[26,27] They use different word frequency-based statistics, not only on title words but also on words and multi-word phrases from entire MedLine records. The original Arrowsmith system only supported a closed discovery approach. Gordon and Lindsay were able to simulate two of Swanson's early discoveries with an open discovery methodology.

Concept-based approaches using UMLS

A different approach was taken up by Weeber and colleagues.[10,12,28] They used the Unified Medical Language System (UMLS)[29] Metathesaurus for identifying biomedically interesting concepts in MedLine titles and abstracts. They also exploited the semantic categorisation that is included. In their simulation of Swanson's early discoveries, for instance, they used a semantic filter based on the categories 'Lipid', 'Vitamin' and 'Element, Ion or Isotope', which represents Swanson's original interest of finding dietary factors that may alleviate his diseases of interest. To identify UMLS concepts, Weeber *et al.* employed MetaMap.[30] MetaMap uses underspecified syntactic analysis to break the text into manageable phrases for further processing. Using the UMLS Specialist Lexicon, it applies extensive variant generation to find the strings in the Metathesaurus containing one or more phrase variants. Also, it uses a linguistically rigorous evaluation metric to determine which Metathesaurus concepts most closely match the original text. MetaMap is also employed in Pratt's[31] and Fuller's[32] literature-based discovery tools. Pratt's LitLinker tool employs the Metathesaurus hierarchy to filter for suitable concepts. Fuller's tool Telemakus extracts concepts from MedLine citations, full text articles and, most notably, analyses figure and table captions to fill a predefined frame based on schema theory. Both Weeber *et al.* and Pratt rank interesting *AC* connections by the number of intermediate *B* concepts.

Visualisation

Van der Eijk and colleagues[33] also use concepts that were extracted from abstracts with commercially available concept recognition software.[34] Once all concepts are found, they use the co-occurrence-based Associative Concept Space (ACS) algorithm to place the extracted concepts in an *n*-dimensional space. This space can be visualised in 2D. Concepts in close proximity are presumably related, and when there is no co-occurrence found among these concepts, a potential discovery has been made. Ranking of potential discoveries is based on the distance between concepts. The Telemakus system[32] also visualises networks in a conceptual graph.

Exploiting MeSH Headings

A different approach is taken by Hristovski *et al.*[19,22] and Srinivasan.[15] They do not use the natural language text from MedLine citations but exploit the manually assigned MeSH indexing terms.[35] While Srinivasan's tool Manjal uses the semantic information that is available for these MeSH terms to successfully replicate all of Swanson's discoveries, Hristovski's tool BITOLA computes Association Rules (ARs) between terms, a technique originating from data-mining research.[36] The results are ranked according to parameters that measure association strength. LitLinker[31] also uses ARs to select (but not rank) interesting associations between concepts.

In the IRIDESCENT discovery tool, Wren *et al.*[17] use a thesaurus of 'objects' that is a combination of different source thesauri focusing on diseases, genes and chemical compounds. They use a fuzzy logic-based algorithm to compute the probability whether two objects are associated. Wren also proposed an extension to the mutual information measure for use in literature-based discovery.[37] The number of intermediate objects is part of the ranking of generated hypotheses.

Concept identification

Up to now, the emphasis of the text-mining techniques employed by LBD tools has been on the correct identification of biomedical terms in text. Most approaches use the UMLS Metathesaurus as the source of terms. When deploying LBD in a genomics context, this may be a limiting factor because the coverage of gene and proteins in the UMLS is low. Recently, Wren *et al.*[17] and Hristovski *et al.*[19] have started to use an extensive thesaurus of gene and protein symbols and names to remedy this. Unambiguously recognising gene symbols, however, is not straightforward as many gene symbols refer to one or more genes or other meanings.[38–40] Wren partly solves this in IRIDESCENT by matching the gene's acronym to its accompanying full name. Literature-based discovery tools should benefit from recent research in identifying gene names[41] and their disambiguation.[42,43]

Information extraction

In previous LBD research, relationships between concepts have been almost exclusively based on co-occurrence statistics and are therefore just mere associations. While concept identification and extraction have been core issues to support LBD, fact extraction has not. This is remarkable because facts such as 'drug *A* lowers blood viscosity' and 'high blood viscosity is a symptom of disease *C*' would render the formulation of a hypothesis rather straightforwardly. The past few years have seen a surge of papers on information extraction, particularly in a genomics context.[44–47] These techniques should be used in the next generation LBD tools to dramatically restrict the number of generated hypotheses.

DISCOVERY SUPPORT TOOLS

As previously mentioned, literature-based discovery is distinctive from more generic text-mining approaches; however, generic text-mining applications might be used for LBD. A fine example is the Chilibot system.[48] Its basic operation is to extract relationships between genes, chemicals and diseases and to visualise these relationships in a network of nodes with edges indicating the type and direction of the relationship. It is possible to look for nodes that are not directly connected but have one (or more) intermediate node(s) that are connected to the disconnected ones. In fact, many more text-mining tools that produce similar biological networks, for instance PubGene,[49] iHOP[50] and Dragon,[51] may be used for LBD purposes. See Hoffmann *et al.*[52] for a recent review of such network text-mining tools. However, the algorithms and user interfaces employed have not been developed to deal with the potential explosion of possible hypotheses by combining the wealth of extracted information. These tools currently have no methods to rank order hypotheses and to provide the user with only the most interesting ones.

Swanson's Arrowsmith tool

Recently, several systems have been developed specifically to support literature-based discovery and hypothesis generation (see Table 1). A brief description is given of five of them that are freely available. An overview of some characteristics of these systems is given in Table 2. Three systems will not be discussed for various reasons. IRIDESCENT, the tool developed by Wren *et al.*[17] has become commercial and is available from eTexx Biopharmaceuticals, Inc. The ACS algorithm and viewer[33] has only limited access. The Telemakus KnowledgeBase System,[32] though freely available, will not be discussed because it has a very focused domain of application, viz. caloric restriction and nutritional aspects of ageing.

Arrowsmith/University of Chicago

Arrowsmith located at University of Chicago is the original Arrowsmith tool developed by Swanson and Smalheiser.[25] The user has to upload two files that contain the results of two PubMed or OVID queries on *A* and *C* subjects, respectively. The server searches for overlapping title words and presents them to the user as the '*B*-List'. The user can edit the *B*-list and view the juxtaposed titles from both literatures for some selected *B*-terms. The user interface is rudimentary, and there is a steep learning curve. Currently, Arrowsmith can be used only for closed discoveries but the next version should also include the possibility of an open discovery approach.

Arrowsmith/University of Illinois at Chicago

This tool is a re-implementation of the original Arrowsmith at Smalheiser's lab at

Table 1: Currently available literature-based discovery systems

System	URL
Arrowsmith/University of Chicago	http://kiwi.uchicago.edu/
Arrowsmith/University of Illinois at Chicago	http://arrowsmith.psych.uic.edu/
BITOLA	http://www.mf.uni-lj.si/bitola/
Manjal	http://sulu.info-science.uiowa.edu/Manjal.html/
LitLinker	http://litlinker.ischool.washington.edu/
ACS	http://www.biosemantics.org/
IRIDESCENT	http://www.etexxbio.com/
Telemakus	http://www.telemakus.net/

Table 2: Freely available literature-based discovery tools and their characteristics in methods and use

Characteristics	Arrowsmith University of Chicago	Arrowsmith University of Illinois at Chicago	BITOLA	Manjal	LitLinker
Registration	No	No	No	Yes	Yes
Online/offline processing	Online	Online	Online	Offline	Online
Concept/terms	Title words	Title words + filtering of UMLS concepts in title words	MeSH and LocusLink	MeSH	UMLS
Documentation	Poor	Poor	Poor	Good	Average
Query formulation	PubMed or OVID (separated from tool)	PubMed (integrated in tool)	Term entry with feedback	Term entry without feedback	Term entry with feedback
Visualisation of results	List of terms, juxtaposition of titles	List of terms, juxtaposition of titles, linkout to PubMed	List of terms, linkout to PubMed	List of terms, linkout to PubMed	List of terms, indication of association strength, title and abstract
User interface	Poor	Average	Average	Average	Advanced
Application domain	General biomedicine	General biomedicine	General biomedicine + focus on genomics	General biomedicine	General biomedicine
Save sessions	Yes	Yes	No	Yes	Yes

the University of Illinois at Chicago.[53] Its major advances are a direct search in PubMed using PubMed's interface, semantic and frequency filtering of concepts, and a more polished user interface. Only closed discoveries are supported.

BITOLA

BITOLA uses an open discovery approach.[19] The user starts with defining a query that is mapped to a concept *X*. Next, the user selects the category of interest, eg diseases, pathologic process. A rank-ordered list of relevant concepts *Y* that are directly related to the query is then presented. Optionally, gene expression localisations can be selected. Next, one or more *Y* concepts are selected together with a target semantic category, eg a drug, and the result is a list of *Z* concepts that are potential discoveries. A linkout to PubMed with an AND query on the *Y* and *Z* concepts is provided to assist human assessment of the potential discovery.

Evaluation of user interface

Manjal

Manjal provides both an open and a closed discovery option.[5] Similar to BITOLA, the user has to select a semantic category of interest. Interestingly, when employing an open discovery process, the final results are automatically computed without having the user to select intermediate concepts. Computation takes some time, therefore the results are not provided online. The user will receive an e-mail message when the results are available. Processing may take some minutes up to several hours, depending on the query and server load.

LitLinker

In LitLinker, an open discovery approach has been implemented.[31] After defining a query, a list of resulting concepts is automatically generated without user intervention of selecting intermediate concepts. The user interface of presenting the results is highly informative and has been evaluated experimentally.[54]

CONCLUSION

Recent advances in literature-based discovery research have resulted both in novel discoveries and usable tools. The domain of most LBD research has been general biomedicine; however, the most recent tools accommodate for more focused genomics discoveries. As text mining in genomics has the specific challenges of gene nomenclature, future tools should closely follow the rapid improvements in the unambiguous identification of genes and proteins mentioned in text. In fact, LBD systems should build upon recent text-mining systems that have been evaluated favourably. The BioCreAtIvE exercise, for instance, has shown that current systems are becoming acceptably robust in extracting gene and protein names from text.[24] Databases filled with facts and relationships extracted using these tools should be the starting point of future LBD systems.

Future literature-based discovery tools should build on recent text mining results

Most user interfaces of text-mining tools considered here are only just adequate to an end user. In many cases, the tools have been developed with specific users in mind who are closely related to the original project. Only recently, a study of user interaction with one tool has been published.[54] For a wider deployment of LBD tools, more research and development are needed to optimally display the discovery results. However, the use of an LBD tool will principally be more complicated than a straightforward literature search as there is no direct evidence of a generated hypothesis. The results are, possibly very diverse, pieces of knowledge that have to be combined and integrated by the users themselves.

Hypotheses that have been formulated with the discussed tools are in many cases highly specific and well motivated. In fact, the better motivated such a potential discovery is, the easier it is to test it. As seen in the LBD literature, novel discoveries are indeed directly testable once animal models or patient data and material are available. We would therefore like to encourage biomedical scientists to use literature-based discovery tools to extend, follow-up, substantiate and explore their ideas, hunches, observations and intuitions.

Acknowledgments

This work was partly supported by the EC funded Network of Excellence INFOBIOMED (IST-2002–507585).

References

1. Langley, P. (2000), 'The computational support of scientific discovery', *Int. J. Human–Comput. Stud.*, Vol. 53, pp. 393–410.
2. Smalheiser, N. R. (2002), 'Informatics and hypothesis-driven research', *EMBO Reports*, Vol. 3(8), p. 702.
3. King, R. D., Whelan, K. E., Jones, F. M. *et al.* (2004), 'Functional genomic hypothesis generation and experimentation by a robot scientist', *Nature*, Vol. 427(6971), pp. 247–252.
4. Hughes, B. W., Moro De Casillas, M. L. and Kaminski, H. J. (2004), 'Pathophysiology of myasthenia gravis', *Semin. Neurol.*, Vol. 24(1), pp. 21–30.
5. Huang, W. X., Huang, P., Fredrikson, S. *et al.* (2000), 'Decreased mRNA expression of TNF-alpha and IL-10 in non-stimulated peripheral blood mononuclear cells in myasthenia gravis', *Eur. J. Neurol.*, Vol. 7(2), pp. 195–202.
6. Matusevicius, D., Navikas, V., Palasik, W. *et al.* (1996), 'Tumor necrosis factor-alpha, lymphotoxin, interleukin (IL)-6, IL-10, IL-12 and perforin mRNA expression in mononuclear cells in response to acetylcholine receptor is augmented in myasthenia gravis', *J. Neuroimmunol.*, Vol. 71(1–2), pp. 191–198.
7. Calabrese, L. and Fleischer, A. B. (2000), 'Thalidomide: Current and potential clinical applications', *Amer. J. Med.*, Vol. 108(6), pp. 487–495.
8. Moller, D. R., Wysocka, M., Greenlee, B. M. *et al.* (1997), 'Inhibition of IL-12 production by thalidomide', *J. Immunol.*, Vol. 159(10), pp. 5157–5161.
9. Moreira, A. L., Sampaio, E. P., Zmuidzinas, A. *et al.* (1993), 'Thalidomide exerts its inhibitory action on tumor necrosis factor alpha by enhancing mRNA degradation', *J. Exp. Med.*, Vol. 177(6), pp. 1675–1680.
10. Weeber, M., Vos, R., Klein, H. *et al.* (2003), 'Generating hypotheses by discovering implicit associations in the literature: A case report of a search for new potential therapeutic uses for thalidomide', *J. Amer. Med. Inform. Assoc.*, Vol. 10(3), pp. 252–259.

11. Swanson, D. R. (1986), ‘Fish oil, Raynaud’s syndrome, and undiscovered public knowledge’, *Perspect. Biol. Med.*, Vol. 30(1), pp. 7–18.

12. Weeber, M., Klein, H., De Jong-van den Berg, L. T. W. and Vos, R. (2001), ‘Using concepts in literature-based discovery: Simulating Swanson’s Raynaud–fish oil and migraine–magnesium discoveries’, *J. Amer. Soc. Inf. Sci. Tech.*, Vol. 52(7), pp. 254–262.

13. Durie, B.G. and Stepan, D. E. (1999), ‘Efficacy of low dose thalidomide (T) in multiple myeloma’, *Blood*, Vol. 94(10, suppl 1, Part 1), p. 316a.

14. Smalheiser, N, R. and Swanson, D. R. (1998), ‘Using ARROWSMITH: A computer-assisted approach to formulating and assessing scientific hypotheses’, *Comput. Methods Programs Biomed.*, Vol. 57(3), pp. 149–153.

15. Srinivasan, P. (2004), ‘Generating hypotheses from MEDLINE’, *J. Amer. Soc. Inf. Sci. Tech.*, Vol. 55(5), pp. 369–413.

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

17. Wren, J. D., Bekeredjian, R., Stewart, J. A. *et al.* (2004), ‘Knowledge discovery by automated identification and ranking of implicit relationships’, *Bioinformatics*, Vol. 20(3), pp. 389–398.

18. Wren, J. D. and Garner, H. R. (2005), ‘Data-mining analysis suggests an epigenetic pathogenesis for Type II diabetes’, *J. Biomed. Biotechnol.*, in press.

19. Hristovski, D., Peterlin, B., Mitchell, J. A. and Humphrey, S. M. (2005), ‘Using literature-based discovery to identify disease candidate genes’, *Int. J. Med. Inform.*, Vol. 74(2–4), pp. 289–298.

20. Cory, K. A. (1997), ‘Discovering hidden analogies in an online humanities database’, *Computers Human.*, Vol. 31(1), pp. 1–12.

21. Gordon, M. D., Lindsay, R. K. and Fan, W. (2002), ‘Literature-based discovery on the World Wide Web’, *ACM Trans Internet Technol.*, Vol. 2(4), pp. 261–275.

22. Hristovski, D., Stare, J., Peterlin, B. and Dzeroski, S. (2001), ‘Supporting discovery in medicine by association rule mining in Medline and UMLS’, *Medinfo*, Vol. 10(Pt 2), pp. 1344–1348.

23. Perez-Iratxeta, C., Bork, P. and Andrade, M. A. (2002), ‘Association of genes to genetically inherited diseases using data mining’, *Nat. Genet.*, Vol. 31(3), pp. 316–319.

24. Blaschke, C., Hirschman, L., Valencia, A. and Yeh, A. (2005), ‘A critical assessment of text mining methods in molecular biology’, *BMC Bioinformatics*, Vol. 6(Suppl 1).

25. Swanson, D. R. and Smalheiser, N. R. (1997), ‘An interactive system for finding complementary literatures: A stimulus to scientific discovery’, *Art. Intell.*, Vol. 91(2), pp. 183–203.

26. Gordon, M. D. and Lindsay, R. K. (1996), ‘Toward discovery support systems: A replication, re-examination, and extension of Swanson’s work on literature-based discovery of a connection between Raynaud’s and fish oil’, *J. Amer. Soc. Inf. Sci. Tech.*, Vol. 47, pp. 116–128.

27. Lindsay, R. K. and Gordon, M. D. (1999), ‘Literature-based discovery by lexical statistics’, *J. Amer. Soc. Inf. Sci. Tech.*, Vol. 50, pp. 574–587.

28. Weeber, M., Klein, H., Aronson, A. R. *et al.* (2000), ‘Text-based discovery in biomedicine: The architecture of the DAD-system’, in ‘Proceedings of the AMIA Symposium’, 4th–8th November, Los Angeles, pp. 903–907.

29. Lindberg, D. A., Humphreys, B. L. and McCray, A. T. (1993), ‘The Unified Medical Language System’, *Methods Inf. Med.*, Vol. 32(4), pp. 281–291.

30. Aronson, A. R. (2001), ‘Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program’, in ‘Proceedings of the AMIA Symposium’, 3rd–7th November, Washington, DC, pp. 17–21.

31. Pratt, W. and Yetisgen-Yildiz, M. (2003), ‘LitLinker: Capturing connections across the biomedical literature’ in ‘International Conference on Knowledge Capture’, ACM Press, New York, pp. 105–112.

32. Fuller, S. S., Revere, D., Bugni, P. F. and Martin, G. M. (2004), ‘A knowledgebase system to enhance scientific discovery: Telemakus’, *Biomed. Digit Libr.*, Vol. 1(1), p. 2.

33. van der Eijk, C. C., van Mulligen, E. M., Kors, J. A. *et al.* (2004), ‘Constructing an associative concept space for literature-based discovery’, *J. Amer. Soc. Inf. Sci. Tech.*, Vol. 55(5), pp. 436–444.

34. van Mulligen, E. M., Diwersy, M., Schmidt, M. *et al.* (2000), ‘Facilitating networks of information’, in ‘Proceedings of the AMIA Symposium’, 4th–8th November, Los Angeles, pp. 868–872.

35. Rogers, F.B. (1963), ‘Medical subject headings’, *Bull. Med. Libr. Assoc.*, Vol. 51, pp. 114–611.

36. Han, J. and Kamber, M. (2000), ‘Data Mining: Concepts and techniques’, Morgan Kaufmann, San Francisco, CA.

37. Wren, J. D. (2004), ‘Extending the mutual information measure to rank inferred literature relationships’, *BMC Bioinformatics*, Vol. 5(1), p. 145.

38. Weeber, M., Schijvenaars, B. J., Van Mulligen, E. M. *et al.* (2003), 'Ambiguity of human gene symbols in LocusLink and MEDLINE: Creating an inventory and a disambiguation test collection', in 'AMIA Annual Symposium Proceedings', 9th–11th November, Washington, DC, pp. 704–708.

39. Tuason, O., Chen, L., Liu, H. *et al.* (2004), 'Biological nomenclatures: A source of lexical knowledge and ambiguity', in 'Proceedings of the 9th Pacific Symposium on Biocomputing, 6th–10th January, Hawaii, pp. 238–249.

40. Chen, L., Liu, H. and Friedman, C. (2005), 'Gene name ambiguity of eukaryotic nomenclatures', *Bioinformatics*, Vol. 21(2), pp. 248–256.

41. Ananiadou, S., Friedman, C. and Tsujii, J. (2004), 'Introduction: Named entity recognition in biomedicine', *J. Biomed. Inform.*, Vol. 37(6), pp. 393–395.

42. Podowski, R. M., Cleary, J. G., Goncharoff, N. T. *et al.* (2004), 'AZuRE, a scalable system for automated term disambiguation of gene and protein names', in 'Computational Systems Bioinformatics (CSB 2004)', Stanford, CA, pp. 415–424.

43. Schijvenaars, B. J., Mons, B., Weeber, M. *et al.* (2005), 'Thesaurus-based disambiguation of gene symbols', *BMC Bioinformatics* (accepted).

44. Blaschke, C., Hirschman, L. and Valencia, A. (2002), 'Information extraction in molecular biology', *Brief. Bioinformatics*, Vol. 3(2), pp. 154–165.

45. Muller, H. M., Kenny, E. E. and Sternberg, P. W. (2004), 'Textpresso: An ontology-based information retrieval and extraction system for biological literature', *PLoS Biol.*, Vol. 2(11), p. e309.

46. Koike, A., Niwa, Y. and Takagi, T. (2005), 'Automatic extraction of gene/protein biological functions from biomedical text', *Bioinformatics*, Vol. 21(7), pp. 1227–1236.

47. Santos, C., Eggle, D. and States, D. J. (2005), 'Wnt pathway curation using automated natural language processing: Combining statistical methods with partial and full parse for knowledge extraction', *Bioinformatics*, Vol. 21(8), pp. 1653–1658.

48. Chen, H. and Sharp, B. M. (2004), 'Content-rich biological network constructed by mining PubMed abstracts', *BMC Bioinformatics*, Vol. 5(1), p. 147.

49. Jenssen, T. K., Laegreid, A., Komorowski, J. and Hovig, E. (2001), 'A literature network of human genes for high-throughput analysis of gene expression', *Nat. Genet.*, Vol. 28(1), pp. 21–28.

50. Hoffmann, R. and Valencia, A. (2004), 'A gene network for navigating the literature', *Nat. Genet.*, Vol. 36(7), p. 664.

51. Pan, H., Zuo, L., Choudhary, V. *et al.* (2004), 'Dragon TF Association Miner: A system for exploring transcription factor associations through text-mining', *Nucleic Acids Res.*, Vol. 32 (Web server issue), pp. W230–234.

52. Hoffmann, R., Krallinger, M., Andres, E. *et al.* (2005), 'Text mining for metabolic pathways, signaling cascades, and protein networks', *Sci. STKE*, Vol. 2005(283), p. pe21.

53. Weeber, M., Torvik, V. I., Swanson, D. R. and Smalheiser, N. R. (2002), 'Enhanced features of the Arrowsmith search engine', in 'Human Brain Project Annual Meeting', Bethedsda, MD.

54. Skeels, M. M., Henning, M., Yetisgen-Yildiz, M. and Pratt, W. (2005), 'Interaction design for literature-based discovery', in 'ACM International Conference on Human Factors in Computing Systems (CHI 2005)', ACM Press, New York.