Advances in Literature-Based Discovery

Marc Weeber

Lister Hill National Center for Biomedical Communications
National Library of Medicine
8600 Rockville Pike
Bethesda, MD 20894
USA
marc@weeber.net

Abstract. Since Swanson's introduction of literature-based discovery in 1986, new hypotheses have been generated by connecting disconnected scientific literatures. In this paper, we present the general discovery model and show how it can be used for drug discovery research. We describe our discovery support tool by discussing a recent discovery for which we used this tool. We conclude by discussing criticisms to and the current status and future of literature-based discovery support tools.

1 Introduction

The amount of scientific knowledge has grown immensely during the past century. Science expands constantly because scientists continue to be curious about the world that surrounds them. If a scientist has found something new, he immediately wonders what its implications are, and tries to formulate new hypotheses that he subsequently tests, which leads to new insights and discoveries. The fact that Nobel prizes, the most prestigious appraisals for scientists, are awarded to people who make breakthrough scientific discoveries, shows that discovery is at the heart of science. Then, the study of *discovery in science*, characterized by Valdés-Pérez as the "generation of novel, interesting, plausible, and intelligible knowledge about the objects of study" [1], is an interesting one. Questions arise as to what the prerequisites are for discovery in terms of existing knowledge and data gathering. How does a scientist recognize patterns in data and how does he define generalizations or even laws? Also, once new facts have been discovered, how does he disseminate and communicate these to other researchers, and how do his colleagues react and integrate this new knowledge? Research into artificial intelligence has tried to analyze and mimic these processes. Some computer systems are able to simulate the discoveries of natural laws based on a database of observations, see [2] for a short overview. Also, computer systems have been developed that assist the human scientist in the scientific discovery process. Both Valdés-Pérez and Langley discuss a wide variety of systems such as MECHEM (catalytic chemistry), ARROWSMITH (biomedicine), GRAFFITI (graph theory), DAVICCAND

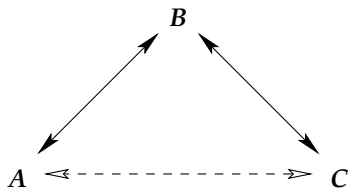

Fig. 1. Swanson's ABC model of discovery. The relationships *AB* and *BC* are known and reported in the literature. The implicit relationship *AC* is a putative new discovery.

(mettalurgy), and MPD/KINSHIP (anthropological linguistics) that have successfully been used to assist in the creation of new scientific knowledge [1, 3].

One of the characteristics of increasing scientific knowledge is that individual scientists have to interpret vast amounts of existing knowledge and acquire specialist skills before they are able to attribute to their scientific domain by discovering new knowledge. Additionally, keeping abreast of the latest developments in order to integrate newly created knowledge with his own research is not a trivial task for a scientist. Simon *et al.* state that scientific publications, as a public blackboard, is the principal instrument for the cumulation and coordination of scientific knowledge [2]. Swanson has shown that it is possible to use scientific publications to generate new knowledge in the context of literature-based discovery.

Our literature-based discovery research has three main goals. First, we integrate Swanson's generic discovery model [4] with Vos's drug discovery model [5]. Second, we use advanced natural language processing (NLP) to efficiently analyze the scientific literature, and third, we develop a tool that may assist researchers in their scientific discovery process. In this paper we will discuss the discovery models, NLP techniques, and the tool in a case study on discovering new applications for the forty year-old drug thalidomide.

2 Models of Discovery

Since 1986, Swanson and his colleague Smalheiser have continuously made discoveries in biomedicine by connecting disconnected knowledge structures, see [6] for an overview. The premise of their approach is that there are two bodies, or structures of scientific knowledge that do not communicate. However, part of the knowledge of one such a domain may complement the knowledge of the other one. Suppose that one scientific community knows that *B* is one of the characteristics of disease *C*. Another scientific group (discipline, or knowledge structure) has found that substance *A* affects *B*. Discovery in this case is making the implicit link *AC* through the *B*-connection. Figure 1 depicts this situation, see also [7].

Vos's model of discovery uses the concept of drug profiles interacting with disease profiles. A profile of a particular drug consists of all the effects it has

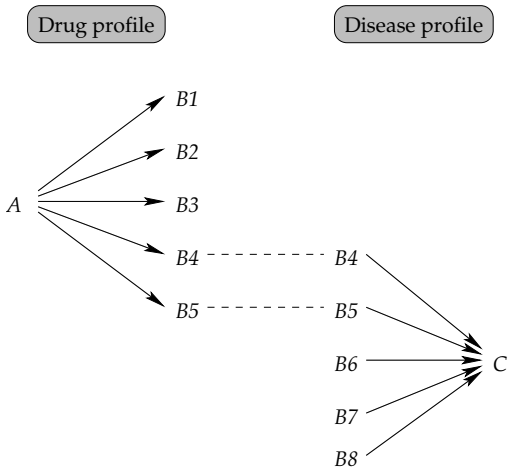

Fig. 2. Vos's and Swanson's model of discovery combined. The linking of a disease profile to a drug profile may be used to find the therapeutic application (disease) *C* for the drug *A* through pathways *B4* and *B5*.

in the human body. Some of them are intended, or *wished for*, i.e. the drug has specifically been developed with these characteristics in mind, others are not wished for. Vos calls all effects the *operational functional characteristics* of a drug. Standard drug development involves the optimization of the wished for characteristics together with a minimization of the negative operational functional characteristics, or adverse effects. However, the not wished for characteristics can be viewed positively in a different context [5, 8, 9]. A well-known example is the anti-hypertensive drug minoxidil. Some patients developed extra hair growth as a not wished for result. Women, for instance, may value this negatively, especially if it concerns facial hair growth. In the different context of baldness, stimulation of hair growth is beneficial. Interestingly, the manufacturers of minoxidil did register male pattern baldness as a new indication for minoxidil. Consequently, hair growth became a new wished for characteristic.

A disease profile consists of a cluster of relevant signs and symptoms, or in other words, the characteristics of the disease. Vos defines the process of drug discovery as the rapprochement of the drug and the disease with respect to their profiles. The more characteristics are relevant to both, the more promising the drug is for treating the disease [5].

Figure 2 shows how Vos's model can be considered as a specification of Swanson's general model in a drug discovery context. The characteristics of the profiles in Vos's model are the intermediate *B*s in Swanson's model. The profile for drug *A*, for instance, may include the therapeutic characteristic (*B*) of "reduction of oxygen demand" whereas "increase of oxygen demand" may be a characteristic of disease *C* [5]. Or, patients with Raynaud's disease (*C*) have

the characteristic of elevated blood viscosity (*B*). One of the characteristics of dietary fish oil (*A*) is blood viscosity reduction [4].

3 Discovery Space

There are two approaches to discovery that we have defined as *open* and *closed* [10]. The closed discovery starts with known *A* and *C*. This may an observed association, or an already generated hypothesis. The discovery in this situation concerns finding novel *B*s that may explain the observation. The open discovery process starts in the knowledge structure in which the scientist takes part (*A*). The first step is to find potential *B*-connections. These will likely be found within the his domain. The crucial step, however, is from *B* to *C* which is most likely outside the scientist's scope, and might therefore be in any point of the knowledge space of science. Or even outside that space. We can illustrate this with the similarity of a person's social life. In a continuously growing world population (total science), our main character (*A*) knows an increasing but limited number of persons (*B*). Keeping up to date with his social structure is not a trivial task for *A*. Knowing the social structure (*C*) of any *B*-person included in his own structure is impossible. Our main character will not know all his friends' friends.

Similar to Swanson, we define discovery as connecting disconnected structures (or disciplines or domains) of scientific knowledge in biomedicine. Note that just any science can be selected, the discovery model holds true for any discovery space. The literature of the selected discipline, biomedicine in our case, is the most comprehensive and accessible format of scientific knowledge in which experimental results, facts, theories, models, and hypotheses are reported. Discovery by connecting different structures implies connecting different (collections of) scientific texts. We therefore pursue *literature-based discovery*. A system that supports literature-based discovery should have the potential of exploring the complete knowledge space. Because we have selected biomedicine as our scientific discipline, we use MEDLINE, the most comprehensive biomedical bibliographical database with over 11,000,000 citations as the representation of the *knowledge universe* in which discoveries may be made. Accessing PubMed [11], the online interface to MEDLINE, and using natural language processing techniques, we have developed a discovery support tool called *Literaby* to explore this vast space.

Literaby implements both the open and the closed discovery approach to discovery. In the open discovery, it first analyzes the literature of the starting point: *A*. Selecting interesting terms, the literature on these *B*-terms is downloaded and analyzed to find the final *C*-term. In the closed discovery, both the literatures on *A* and *C* are downloaded and analyzed to search for interesting overlapping *B*-terms to strengthen (or reject) the initial *AC*-hypothesis. In most cases, an open discovery concerns *generating* a hypothesis that is *evaluated* in a closed discovery process.

4 Text Analysis

Swanson's first discovery of the probable therapeutic effects of fish oil on patients with Raynaud's disease [4] was a coincidence (Swanson, personal communication). He was asked to study the literature on the Inuit diet. Fish is a main ingredient of this diet, and the effects of fish oil on the the cardiovascular system in Inuit has been studied. Reduced blood viscosity and blood platelet aggregation, and certain vasoreactive characteristics were observed in Inuit. In another context, Swanson had been studying the literature on Raynaud's disease. From this literature he had learned that patients with this disease have a relatively high blood viscosity and increased platelet aggregation function. Also, they were characterized by certain vasoreactive phenomena. Combining this knowledge, he hypothesized that the active ingredients of fish oil, omega-3 fatty acids, may help Raynaud's patients. With this hypothesis in mind, he studied the literatures both on fish oil and on Raynaud's disease to find out that there was no overlap at that time (1986). Using the model of disconnected bodies of biomedical knowledge, he published a second hypothesis that magnesium insufficiency is involved in migraine. No one had pointed this out in the literature, while Swanson found eleven indirect connections in the literature [12].

The first two discoveries were done by extensive manual searching in literature databases and reading many titles and abstracts of scientific publications. Since 1988, Swanson has used computational text analysis tools to assist him in studying the literature. These tools have evolved into a discovery support tool called ARROWSMITH [7]. The user can upload a file of titles on A and on C (an implementation of the closed approach). The tool provides a list of overlapping Bs. Additionally, the context of the Bs can be viewed in a juxtaposed (AB next to BC-sentences). The list of B-terms is potentially very long, and filtering is needed. The current analytic approach is to use an extensive stop list, a list of words such as determiners and adverbs that are considered non-relevant. This list has mainly been compiled during rediscovering his first discoveries, incorporating expert knowledge from users.

Gordon and Lindsay used a more principled analytic approach based on word frequency (lexical) statistics used in Information Retrieval research [13, 14]. They are mainly interested in the open discovery approach. They are able to replicate Swanson's first two discoveries. In [13] they use specific measures and provide a likely explanation why these techniques work in the Raynaud–fish oil case. However, when applied to the migraine–magnesium case, the same statistics fail and different ones had to be used [14]. Therefore, there is still not a unifying, principled lexical statistical approach.

Our approach to the analysis of titles and abstracts of scientific publications is to use advanced NLP techniques to identify biomedical concepts in text. The Unified Medical Language System (UMLS)® [15] provides the largest biomedical thesaurus to date: the Metathesaurus®. The Metathesaurus provides a uniform, integrated distribution format from over 60 biomedical source vocabularies and classifications, and links many different names for the same con-

cept. Over 700,000 biomedical concepts are represented with over 1,500,000 text strings.

The use of concepts has several advantages. First, different textual representations, i.e. spelling variants, synonyms, derivations, and inflections are all linked to one concept. For instance, IL-12, IL12, interleukin 12, CLMF, cytotoxic lymphocyte maturation factor(s), and natural killer cell stimulatory factor(s) refer to the same concept: *Interleukin-12*. Second, many biomedical ideas or concepts are expressed by more than one word. Finding meaningful multi-word terms in text is non-trivial in NLP. Different word statistical strategies may be employed [16], and results always include noise. By using concepts, we select only existing, biologically meaningful, ones. We employ the MetaMap program [17, 18, 19] to find UMLS concepts in natural language text.

The most important reason to use concepts, however, is the availability of the UMLS semantic classification scheme. Each concept has been assigned to one or more semantic categories. There is a total of 134 categories including "Disease or Syndrome", "Gene or Genome", "Amino Acid, Peptide, or Protein". The concept *Thalidomide*, for instance, has been assigned the semantic types "Organic Chemical", "Pharmacologic Substance", and "Hazardous or Poisonous Substance". At different stages of the discovery process, we can select only certain semantic types to filter the output of the text analysis. For instance, if we are looking for diseases in text, we select only the semantic type "Disease or Syndrome" which will result in a list of disease concepts extracted from natural language sentences. Figure 4 provides a part of the interface to semantic filter in our discovery system. We provide a more extensive overview of the our text analysis techniques in [10].

5 Literaby and Thalidomide

Literaby, our current, web-based, discovery support tool has evolved from our first tool [20] with respect to the query generation phase (now fully automated) and the interface that presents the bibliographic evidence to the user. By following the discovery of new therapeutic applications for the drug thalidomide [21], we will show how Literaby assists scientists in literature-based discovery.

Between 1959 and 1961, thalidomide was a popular over the counter sedative. Devastating teratogenic effects led to withdrawal from the market only a few years after its introduction. In recent years, however, interest in thalidomide has intensified based on its reported anti-inflammatory and immunomodulatory properties. In 1998, the FDA approved thalidomide for the indication of erythema nodosum leprosum, an inflammatory manifestation of leprosy. Additionally, thalidomide seems to have beneficial effects on ulcers and wasting associated with HIV infection.

The first step (*A* in the discovery model) is to identify concepts in the UMLS that are related to thalidomide. Entering the string `thalidomide` results in a list of 33 concepts that map to this string. Figure 3 depicts part of this list.

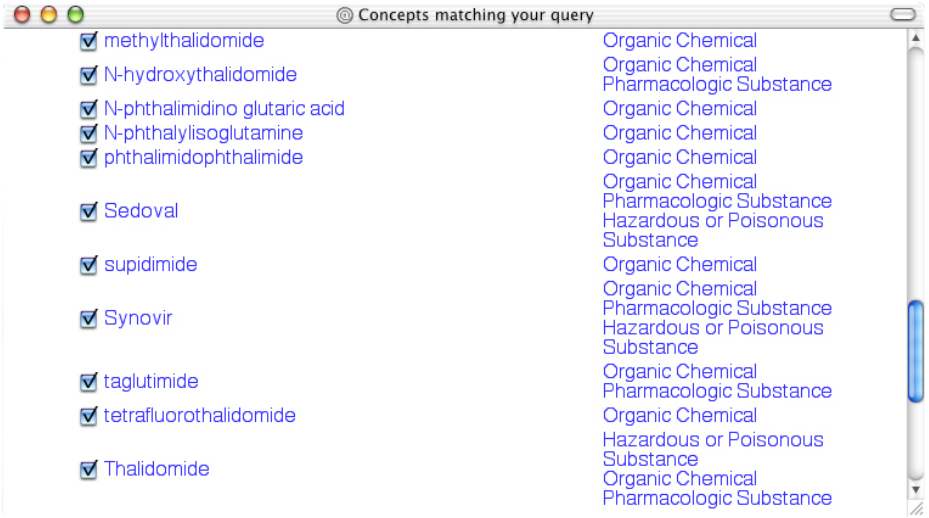

Fig. 3. The search `thalidomide` in the UMLS Metathesaurus resulted in 33 concepts, for instance, different chemical names for the substance, but also brand names for the drug.

By using the hierarchy of the thesaurus we not only find the concept *Thalidomide*, which is the generic name of the drug, but also the brand names, which are children concepts in the thesaurus, and the chemical description of the compound. The user has the option to (de)select these concepts, and may proceed. Employing MetaMap, Literaby maps the concepts back to their textual variants to automatically generate and execute a query to PubMed. The resulting citations are downloaded and analyzed to extract concepts. After this step, the user is involved again; the *B*-concepts have to be selected. For this, the user's expert knowledge is needed. In this case, we collaborated with an immunologist, because the newly registered application involves the immune system. We hypothesized that we might find new therapeutic applications through thalidomide's apparently successful immunologic pathways. Literaby presents the user the semantic filter, i.e., the list of the 134 categories or semantic types (Fig. 4).

At this stage, we select only the semantic type of "Immunologic Factor", and Literaby returns a list of 93 immunologic factors that co-occur in sentences mentioning a textual representation of the concept *Thalidomide*. Figure 5 shows the twelve most frequent ones. The domain expert selected *Interleukin-12* (IL-12) as the *B*-concept of potential interest. Clicking on the button before the concept, the user may see the sentences in which this *B*-concept co-occurs with thalidomide. For *Interleukin-12*, we observe sentences such as:

– Inhibition of **IL-12** production by *thalidomide*.

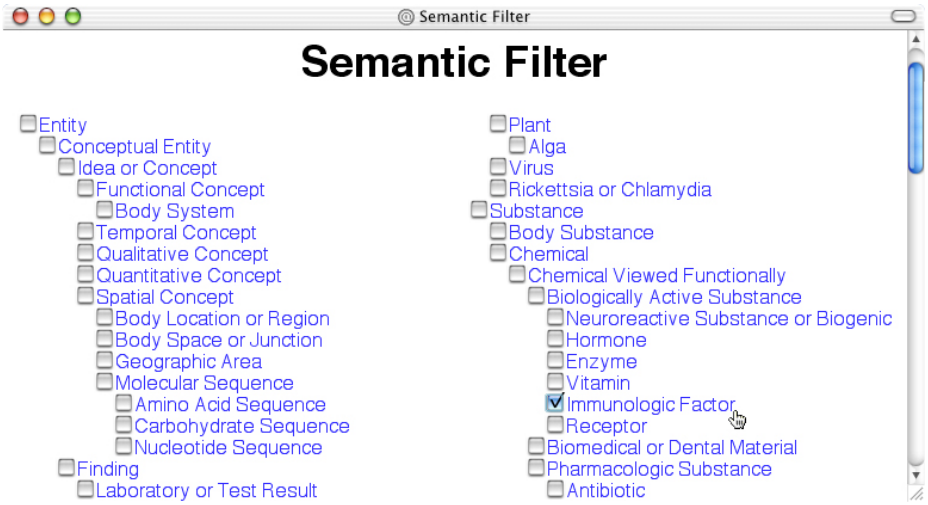

Fig. 4. The semantic filter of the discovery support tool Literaby. There are 134 semantic types that the user may select.

– *Thalidomide* potentially suppressed the production of **IL-12** by PBMC [...].
– *Thalidomide*-induced inhibition of **IL-12** production was additive [...].

It appears that thalidomide has inhibitory effects on IL-12. However, there is also some bibliographical counter-evidence:

– *Thalidomide* stimulates [...] **IL-12** production in HIV patients.

IL-12 inhibition, together with the reported stimulatory effect on IL-10 production, seems to be the mechanism of how thalidomide favors the differentiation of T-helper 0 (Th0) immune system cells into T-helper 2 (Th2) cells by blocking differentiation of Th1 cells. Our hypothetical model of action [21] suggests that patients with, in particular, auto-immune diseases may benefit from thalidomide treatment.

Using *Interleukin-12* as the selected *B*-concept, we downloaded all citations from PubMed that include (variants of) IL-12 in either title or abstract. The resulting citations were MetaMapped to UMLS concepts, and Literaby provides the user again with the semantic filter. At this stage, we looked for *C*-concepts, disease concepts in our case. We selected therefore the semantic type "Disease or Syndrome", which resulted in a list of 420 diseases that co-occur with IL-12. After a partly automated filtering process, see [21], we studied the sentences that related IL-12 to the reduced set of diseases. Examples are:

– **IL-12** [...] expression in mononuclear cells in response to acetylcholine receptor is augmented in *myasthenia gravis*.

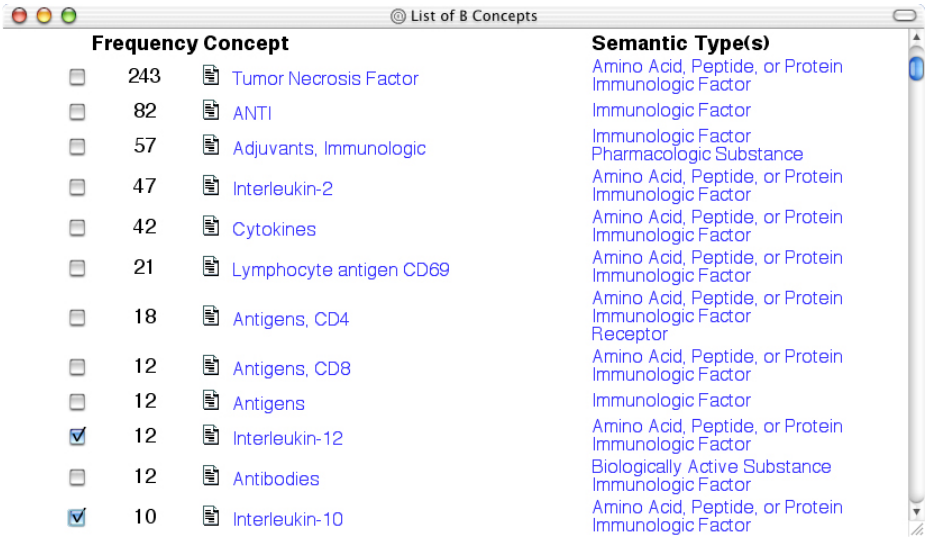

Fig. 5. Top of the list of immunologic factors that co-occur in sentences with the *Thalidomide*.

- Possible involvement of **IL-12** expression by Epstein-Barr virus in *Sjögren syndrome*.
- *Acute pancreatitis* patients had serum concentrations of total **IL-12, IL-12p40**, and IL-6 significantly higher ($p < 0.05$) than those of the healthy subjects.
- Expression of B7-1, B7-2, and **IL-12** in anti-Fas antibody-induced *pulmonary fibrosis* in mice.

The previous sentences indicate that IL-12 is overexpressed in these diseases. Studying the sentences, their complete abstracts, and sometimes even the online full text papers, we hypothesized for twelve diseases that thalidomide might be a useful therapy through its inhibitory effects of IL-12. These twelve hypotheses were the starting point of twelve closed discovery processes Literaby downloaded and analyzed each of these *C*-literatures. The discovery process consisted of finding (a lack of) overlapping immunologic *B*-concepts to strengthen (or reject) the initial hypotheses.

Using chronic hepatitis C (CHC) as an example, the semantic filter, again set to "Immunologic Factor", provided us with a list of 60 immunologic factors, presented in a similar way as Fig. 5. We find additional citations in the CHC literature that IL-12 is augmented in patients with this disease. Figure 6 provides the interface in which the bibliographical information on thalidomide–IL-12 and IL-12–CHC is juxtaposed. In one overview the user can assess the *AB* and *BC*-information to infer the hypothesis *AC*.

In addition to IL-12, we also find the concept *Tumor Necrosis Factor* (TNFα). It is widely known that thalidomide inhibits TNFα through mRNA degradation.

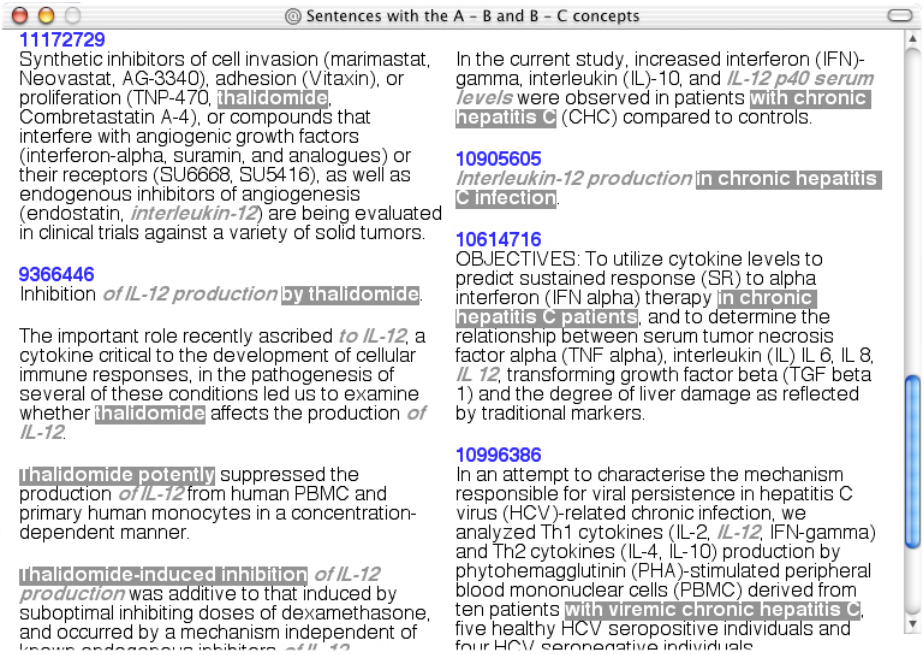

Fig. 6. Bibliographic information that suggests that chronic hepatitis C may benefit from thalidomide through IL-12 inhibition. The left panel shows sentences in which *A* and *B*-concepts co-occur, the right panel shows the relevant sentences for *B* and *C*.

It turns out that CHC is characterized by increased levels of TNFα. Thus, we have strengthened our initial hypothesis that thalidomide may be used in CHC by elucidating an additional pathway.

In the closed discovery processes, we were able to find strong bibliographical evidence that supports the hypotheses that thalidomide may be a therapeutic drug for helicobacter pylori-induced gastritis, acute pancreatitis, chronic hepatitis C, and myasthenia gravis. For the latter three serious diseases, there is no known cure or therapy. The bibliographical findings merit experimental and clinical studies that should provide information on the cost/benefit trade-off of effects and side effects of thalidomide in these diseases.

6 Discussion

In the presented example, the discovery was made by human scientists supported by a tool for analyzing huge amounts of text. We do not regard, or pursue, literature-based discovery as an automatic process. The reason for this is that expert knowledge is indispensable in studying the output of the support system, not only to filter out non-interesting information but also to assess po-

tentially contradicting information. For instance, there is one MEDLINE citation that co-mentions thalidomide and myasthenia gravis and claims that that thalidomide is not effective in Lewis rats that suffer from myasthenia gravis. This information potentially refutes our hypothesis that thalidomide may be benefical for patients with this disease, however, the expert provided the knowledge that Lewis rats have an altered immune system. Conclusions based on these experiments may therefore not be transferred to a human context. We think it impossible to model such domain knowledge in a discovery system. Even if it is possible to model knowledge to such detailed extent, one has to consider that the model should comprise the total biomedical knowledge available, because this is knowledge space in which literature-based discovery takes place. The second reason why we do not pursue automated discovery is that it will result in just another database, in this case one of hypotheses. How to make a decision as to what hypothesis to test experimentally? Again, human experts are needed to decide. Some bibliographically well founded hypotheses may not interesting to test. For instance, thalidomide has some severe side effects, a clinical application may therefore be only interesting in severe diseases or diseases for which there is no treatment at all.

There is some scepticism towards literature-based discovery and its potential for scientific research. Results are considered too obvious and once an hypothesis is proposed, people might say "it's logical" or "of course", and the hypothesis may have activated existing knowledge that was already available in one person. We have also encountered remarks such as "but then you can also hypothesize that…" with the intend to downplay the discovery, but actually with the result of generating yet another plausible, partly literature-based, discovery. We can counter these criticisms with two facts. First, Swanson and his colleague Smalheiser have made eight literature-based discoveries that have been published in relevant, peer-reviewed, scientific journals. Swanson's first two discoveries have even been corroborated experimentally and clinically. The findings for our drug discovery is currently in submission, and if it will be accepted, we can assume that scientists value our contribution as interesting.

Second, no one has denied the premise of the model, viz. that there are disconnected structures in science that may benefit from connection. This is shown by the relative ease with which we have discovered new hypothetical applications for the controversial and well-known drug thalidomide. This is not surprising, because biomedical scientists work in widely varying and highly specialized disciplines and contexts. For instance, we observe a distinction between *in vivo*, or clinical research in humans, *in vitro*, preclinical research in laboratory and animal experiments, and *in silico*, computer-based research. The transfer of knowledge from one domain to the other is non-trivial. The research interests and goals of both domains are very different. Also, educational background of the scientists diverges largely, being clinical (medicine), experimental (biology, pharmacy, biochemistry), or computational (computer science, mathematics), respectively.

Current literature-based discoveries have mainly been made in biomedicine. Both Swanson and Spasser [22] have noted that the biomedical bibliography is particularly suited for this because of the explicit titles that often state the main outcome of the research, for instance:

- Inhibition of IL-12 production by thalidomide.
- Thalidomide treatment in chronic constrictive neuropathy decreases endoneurial TNFα, increases IL-10 and has long-term effects on spinal cord dorsal horn met-enkephalin.
- Inhibition of TNFα synthesis with thalidomide for prevention of acute exacerbations and altering the natural history of multiple sclerosis.

However, not only titles are interesting. In the thalidomide case, there are only two titles mentioning IL-12 together with the drug. There were ten more sentences in MEDLINE abstracts that provided additional useful information. Of course, using abstracts also introduces more noise, but the employed filtering techniques were able to suppress this. More importantly, Cory showed that literature-based discovery is possible in humanities, a scientific discipline that is not famous for its explicit titles [23]. This suggests that the presented approach to generating scientific hypotheses is valid for science in general. As long as there are comprehensive bibliographic databases, reported knowledge can be combined to generate new, hypothetical knowledge. Additionally, it would be interesting to combine databases from different disciplines. Biomedicine may profit from more chemically and biologically oriented databases, such as Biological and Chemical Abstracts. But even wider gaps between disciplines may result in interesting new insights.

Research in literature-based discovery has been acknowledged its importance in information and library sciences but unfortunately, it has received little attention in biomedicine. In seems that the disconnection between biomedicine and information science prevents further developments and use of the ideas of Swanson [22]. Recently, however, an NIH grant has been awarded to Dr. Smalheiser (University of Illinois at Chicago) in the context of The Human Brain Project and neuroinformatics (Smalheiser, personal communication). The goal of this project is to use informatics tools to optimize communication between neuroscientists and to connect individual research projects, data, and results. ARROWSMITH will be developed further and used as one of the tools to reach this goal. This research is the first step in transferring literature-based discovery support tools from the computer and information science lab into the wet lab.

Acknowledgments

Over the past years, the presented research has benefited from the input of many people. I would like to thank Rein Vos, Lolkje de Jong - van den Berg, Henny Klein, and Don Swanson for their many contributions and discussions and Grietje Molema as the domain expert in the thalidomide discovery. I am

grateful to Alan Aronson and Jim Mork for access to the National Library of Medicine's natural language processing tools.

This research was supported in part by an appointment to the National Library of Medicine Research Participation Program administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the National Library of Medicine.

References

[1] Raúl E. Valdés-Pérez. Principles of human computer collaboration for knowledge discovery in science. *Artificial Intelligence*, 107(2):335–346, 1999.

[2] Herbert A. Simon, Raúl E. Valdés-Pérez, and Derek H. Sleeman. Scientific discovery and simplicity of method. *Artificial Intelligence*, 91(2):177–181, 1997.

[3] Pat Langley. The computational support of scientific discovery. *International Journal of Human-Computer Studies*, 53:393–410, 2000.

[4] Don R. Swanson. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, 30(1):7–18, 1986.

[5] Rein Vos. *Drugs looking for diseases*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1991.

[6] Neil R. Smalheiser and Don R. Swanson. Using ARROWSMITH: A computer-assisted approach to formulating and assessing scientific hypotheses. *Computer Methods and Programs in Biomedicine*, 57:149–153, 1998.

[7] Don R. Swanson and Neil R. Smalheiser. An interactive system for finding complementary literatures: A stimulus to scientific discovery. *Artificial Intelligence*, 91(2):183–203, 1997. `http://kiwi.uchicago.edu`.

[8] Floor Rikken and Rein Vos. How adverse drug reactions can play a role in innovative drug research. *Pharmacy World & Science*, 17(6):195–200, 1995.

[9] Floor Rikken. *Adverse drug reactions in a different context*. PhD thesis, University of Groningen, The Netherlands, 1998.

[10] Marc Weeber, Rein Vos, Henny Klein, and Lolkje T. W. de Jong-van den Berg. Using concepts in literature-based discovery: Simulating Swanson's Raynaud – fish oil and migraine – magnesium discoveries. *Journal of the American Society for Information Science and Technology*, 52(8):548–557, 2001.

[11] National Library of Medicine. PubMed. `http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed`, 2001.

[12] Don R. Swanson. Migraine and magnesium: Eleven neglected connections. *Perspectives in Biology and Medicine*, 31(4):526–557, 1988.

[13] Michael D. Gordon and Robert K. Lindsay. 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. *Journal of the American Society for Information Science*, 47(2):116–128, 1996.

[14] Robert K. Lindsay and Michael D. Gordon. Literature-based discovery by lexical statistics. *Journal of the American Society for Information Science*, 50(7):574–587, 1999.

[15] Donald A. B. Lindberg, Betsy L. Humphreys, and Alexa T. McCray. The unified medical language system. *Methods of Information in Medicine*, 32(4):281–291, 1993.

[16] Marc Weeber, Rein Vos, and R. Harald Baayen. Extracting the lowest-frequency words: Pitfalls and possibilities. *Computational Linguistics*, 26(3):301–317, 2000.

[17] Thomas C. Rindflesch and Alan R. Aronson. Ambiguity resolution while mapping free text to the UMLS Metathesaurus. In *Proceedings of the Eighteenth Annual Symposium on Computer Applications in Medical Care*, pages 240–244. Hanley and Belfus, Philadelphia, PA, 1994.

[18] Alan R. Aronson. The effect of textual variation on concept based information retrieval. In *Proceedings of the 1996 AMIA Annual Fall Symposium*, pages 373–377. Hanley and Belfus, Philadelphia, PA, 1996.

[19] Alan R. Aronson. Effective mapping of biomedical text to the umls metathesaurus: The MetaMap program. In *Proceedings of the 2001 AMIA Annual Fall Symposium*, page submitted. Hanley and Belfus, Philadelphia, PA, 2001.

[20] Marc Weeber, Henny Klein, Alan R. Aronson, James G. Mork, Lolkje T. W. de Jong-van den Berg, and Rein Vos. Text-based discovery in biomedicine: The architecture of the *DAD*-system. In *Proceedings of the 2000 AMIA Annual Fall Symposium*, pages 903–907. Hanley and Belfus, Philadelphia, PA, 2000.

[21] Marc Weeber, Rein Vos, Henny Klein, Lolkje T. W. de Jong-van den Berg, and Grietje Molema. Discovering new knowledge in the biomedical literature using a computer support tool: Four hypothetical therapeutic applications for thalidomide. *submitted*, 2001.

[22] Mark A. Spasser. The enacted fate of undiscovered public knowledge. *Journal of the American Society for Information Science*, 48(8):707–717, 1997.

[23] Kenneth A. Cory. Discovering hidden analogies in an online humanities database. *Computers and the Humanities*, 31:1–12, 1997.