Drug Discovery as an Example of Literature-Based Discovery

Marc Weeber*

Lister Hill National Center for Biomedical Communications
National Library of Medicine, Bethesda, Maryland, 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 have developed a discovery support tool that employs Natural Language Processing techniques to extract biomedical concepts from Medline titles and abstracts. Using semantic knowledge, the user, typically a biomedical scientist, can efficiently filter out irrelevant information. This chapter provides an algorithmic description of the system and presents a potential drug discovery. We conclude by discussing the current and future status of literature-based discovery in the biomedical research domain.

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.

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" (Valdés-Pérez, 1999), 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

* The author can be contacted at the Dept. of Medical Informatics, Erasmus University Rotterdam, Dr. Molewaterplein 50, 3015 GE Rotterdam, The Netherlands.

S. Džeroski and L. Todorovski (Eds.): Computational Discovery, LNAI 4660, pp. 290–306, 2007.

laws based on a database of observations, see (Simon et al., 1997) 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 (mettalurgy), and MPD/KINSHIP (anthropological linguistics) that have successfully been used to assist in the creation of new scientific knowledge (Valdés-Pérez, 1999; Langley, 2000).

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 contribute 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. (1997) state that scientific publications, as a public blackboard, is the principal instrument for the cumulation and coordination of scientific knowledge. Swanson has shown that it is possible to use scientific publications to generate new knowledge in the context of literature-based discovery.

This chapter describes our reseach in literature-based discovery (Weeber, 2001). Our goals are three-fold. First, we integrate Swanson's generic discovery model (Swanson, 1986) with Vos's drug discovery model (Vos, 1991). 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 (Smalheiser & Swanson, 1998) 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 (Swanson & Smalheiser, 1997).

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 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.

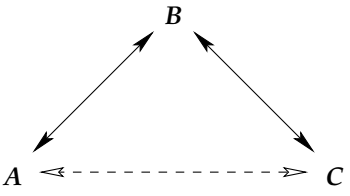

Fig. 1. Swanson's ABC model of discovery. A, B, and C are general concepts. The relationships AB and BC are known and reported in the literature. The implicit relationship AC is a putative new discovery.

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 (Vos, 1991; Rikken & Vos, 1995; Rikken, 1998).

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 (Vos, 1991).

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 Bs 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 (Vos, 1991). 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 (Swanson, 1986).

3 Discovery Space

There are two approaches to discovery that we have named as *open* and *closed* (Weeber et al., 2001). The closed discovery starts with known A and C. This may be an observed association, or an already generated hypothesis. The discovery in this situation concerns finding novel Bs that may explain the observation. In the model, the letters A, B, and C refer to general scientific concepts that researches use.

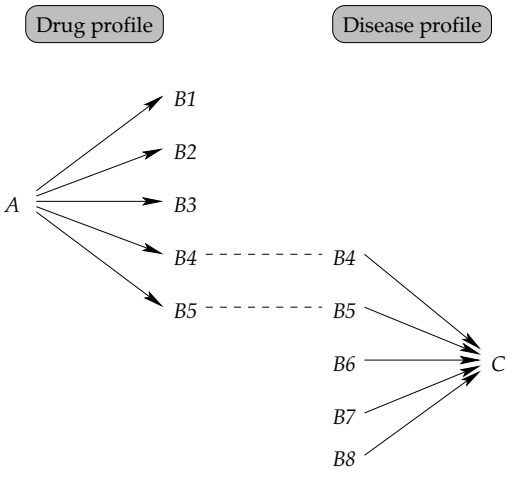

Fig. 2. Vos's and Swanson's model of discovery combined. The linking of a disease profile to a drug profile may be used to find the therapeutic application (disease) C for the drug A through pathways (therapeutic characteristics) *B4* and *B5*.

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 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.

A closed discovery process starts with an initial hypothesis that A has some association to C. The nature of this association is unknown or not fully understood. The goal of the closed approach is first to unveil new possible explanations for an AC association, and second to provide an evaluation of the strength of the association. The likely outcome of the closed approach is to either strengthen or to reject the AC association.

Similar to Swanson, we define discovery in biomedicine as connecting disconnected structures (or disciplines or domains) of biomedical 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. Each citation consists of at least a title. In many cases, an abstract is available as well. Also, other bibliographic information is included, such as authors, journal, date of publication, and keywords (called Medical Subject Headings, or MeSH). Using PubMed (`http://www.pubmed.gov`), the online interface to MEDLINE, and using NLP techniques, we have developed a discovery support tool called *Literaby* to explore this vast space.

The definition of the discovery space allows us to specify the model letters A, B, and C used in the model. Swanson uses MEDLINE titles, therefore, the model letters are (combinations of) title words. As our implementation of the model is concept based, the letters are represented by biomedical concepts (see next section).

Literaby implements both the open and the closed 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 process.

4 Text Analysis

Swanson's first discovery of the probable therapeutic effects of fish oil on patients with Raynaud's disease (Swanson, 1986) 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 the knowledge from two contexts, 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 (Swanson, 1988).

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 (Swanson & Smalheiser, 1997).

In ARROWSMITH, the user can upload a file of Medline 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. Also, words with a too general biomedical meaning are included in this list. The stop 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 (IR) research (Gordon & Lindsay, 1996; Lindsay & Gordon, 1999). In addition to MEDLINE titles, they use the abstracts of citations as well. Gordon and Lindsay emply the statistics to find a rank-ordered list of potentially relevant words. They use the most highly ranked words to walk through their open discovery approach.

Gordon and Lindsay are able to replicate Swanson's first two discoveries. Gordon and Lindsay (1996) 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 (Lindsay & Gordon, 1999). Therefore, there still does not exist 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)® (Lindberg et al., 1993) 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 concept. 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 (Weeber et al., 2000b), and results always include noise. By using concepts, we select only existing, biologically meaningful, ones. We employ the MetaMap program (Rindflesch & Aronson, 1994; Aronson, 1996; Aronson, 2001) 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 on page 300 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 (Weeber et al., 2001).

Hristovski et al. (2001) also use a concept-based approach. They use MeSH keywords added to Medline citations. The UMLS provides co-occurrence tables of major MeSH keywords. Using these frequency data, Hristovski et al. (2001) apply association rules to compute the most interesting associations. In an interactive interface, the user, typically a biomedical scientist, can quickly assess these associations. In this book, Hristovski and his collegues show how their system can be used.

5 Literaby

Literaby, our current, web-based, discovery support tool has evolved from our first tool called the *DAD*-system (Weeber et al., 2000a). The acronym *DAD* expands to Disease – Adverse Drug Reaction – Drug, or the other way around. It represents our interest in drug discovery. The new version, Literaby, shows that our analytic approach can be generalized. Other changes are that the query generation phase is now fully automated, and the interface for presenting the bibliographic evidence has been overhauled.

The underpinnings, however, have not been changed. This section provides an algorithmic overview of the semi-automated discovery process. The high level

Table 1. High level description of the Literaby system

GIVEN: Current version of Medline, Metathesaurus
INPUT: text string A
1. *open* discovery phase: generating a hypothesis
a. From A to B using the Algorithm 1 (automatic)
b. User selects most promising B-concepts based on computer output and literature
c. From B to C using Algorithm 2 (automatic)
d. User selects most promising C-concepts:these are the putative new discoveries
2. *closed* discovey phase: evaluating a hypothesis (Algorithm 2)
OUTPUT: Set of concepts C that have likely new connections to A through B

Table 2. Literaby Algorithm 1: one step in the *open* process

GIVEN: Current version of Medline, Metathesaurus
INPUT: text string A
1. Find set of concepts SA for text string A in Metathesaurus
2. Find textual variants SA-VAR for SA using MetaMap
3. Find titles and abstracts A-CIT in Medline query composed of SA-VAR
4. Find ALL-B concepts in titles and abstracts of A-CIT using MetaMap
5. Select set of concepts SEN-B from ALL-B that co-occur in sentences with SA
6. User forms a semantic filter FILT-B by subsetting the 134 semantic categories
7. Apply semantic filter FILT-B to SEN-B to retrieve set of concepts SB
OUTPUT: Set of concepts SB

Table 3. Literaby Algorithm 2: *closed* process

GIVEN: Current version of Medline, Metathesaurus
INPUT: text string A, text string C
1. Find set of concepts SA for A, SC for C in Metathesaurus
2. Find textual variants SA-VAR for SA and SC-VAR for SC using MetaMap
3. Find citations A-CIT and C-CIT in Medline using SA-VAR and SC-VAR, respectively
4. Find ALL-AB concepts in titles and abstracts of A-CIT using MetaMap
5. Find ALL-BC concepts in titles and abstracts of C-CIT using MetaMap
6. Select set of concepts SUB-AB from ALL-AB that co-occur in sentences with SA
7. Select set of concepts SUB-BC from ALL-BC that co-occur in sentences with SC
8. Select potential concepts POT-B such that SUB-AB = SUB-BC
9. User forms a semantic filter FILT-B by subsetting the 134 semantic categories
10. Apply semantic filter FILT-B to POT-B to retrieve set of concepts SB
OUTPUT: Set of concepts SB

description is presented Table 1. The user starts with an *open* discovery process. He starts with his term of interest, for instance, a drug. The system then tries to find concepts that are in some way related to this drug using algorithm 1 (Table 2).

Algorithm 1 maps the initial query string to biomedical concepts, and "back-translates" this concept to all synonyms and textual varians that are available in the natural language text of Medline. Literaby formulates the query to PubMed, the online version of Medline and retrieves the citations that matched the query. Literaby also takes care of finding concepts in these citations through MetaMap, and finally assists the user in selecting a semantic filter (or use a predefined one).

The output of Algorithm 1 is a set of concepts that have some (user defined semantic) relationship with the starting text string. In case of a drug, typical

B-concepts are processes that are a likely biologic actions of this drug. The user selects a few most promising ones, basing his selections on expert knowledge, and assessing the bibliographic evidence provided by Literaby.

The next step is a replay of Algorithm 1, but now with the B-concepts as input, and likely C-concepts as output. Typical C-concepts in the case of drug discovery are diseases or pathological processes. Again using the available bibliographic information, the user selects the most interesting C-concepts to start the second phase of the discovery process, the CLOSED discovery using Algorithm 2 (Table 3).

With this algorithm, the user tries to evaluate the generated hypotheses in the open discovery process. The main idea is that the more relations (Bs) there are between A and C, the more plausible the association AC is. In the next section, we illustrate the use of Literaby to assist scientists by following the discovery of new potential therapeutic applications for the drug thalidomide (Weeber et al., 2003).

6 Literaby and Thalidomide

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.

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 then proceeds. Employing MetaMap, Literaby maps the concepts back to their textual variants to automatically generate and execute a query to PubMed. For instance, the text string `thalidomide` maps to the concept *Thalidomide*. The UMLS provides us the drug brand name, among other "thalidomid", "supidimide" and "sedoval". These brand names are included in the PubMed query.

The resulting citations are downloaded and analyzed to extract concepts from the titles and abstracts, if available. 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 pathway modulation. Literaby presents the semantic filter

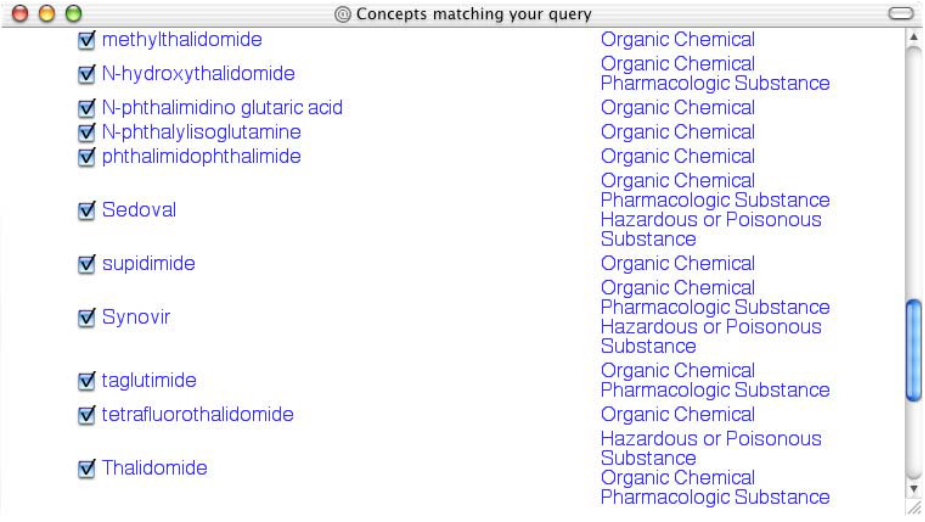

Fig. 3. The search string `thalidomide` in the UMLS Metathesaurus resulted in 33 concepts, for instance, different chemical names for the substance, but also brand names for the drug

to the user, where he can choose from a the list of the 134 categories or semantic types (Fig. 4).

At this stage, we only select 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*.
- *Thalidomide* potentially suppressed the production of **IL-12** by PBMC [...].
- *Thalidomide*-induced inhibition of **IL-12** production [...].

Indeed, it appears that thalidomide's inhibitory effects on IL-12, 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 (Weeber et al., 2003) 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,

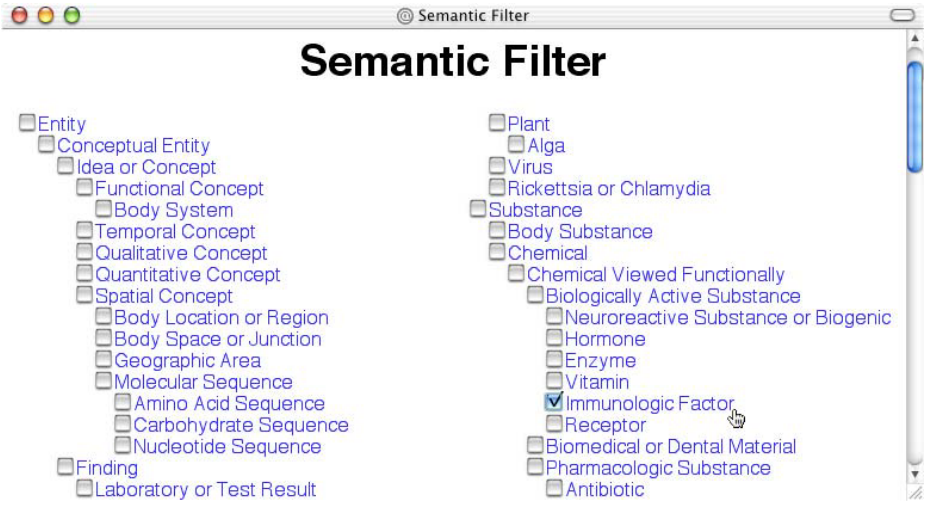

Fig. 4. The semantic filter of the discovery support tool Literaby. There are 134 semantic types that the user may select.

disease concepts in our case. We selected 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 Weeber et al. 2003), 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*.
- 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

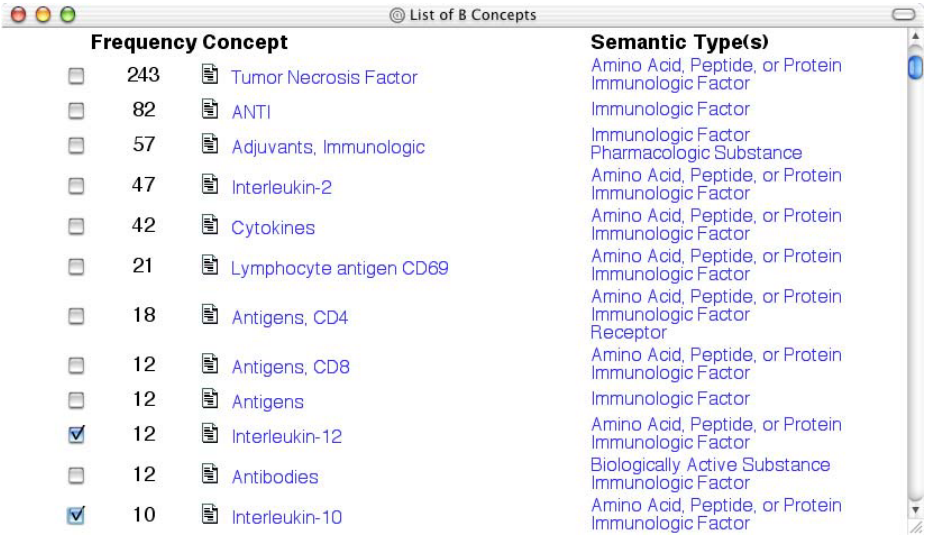

Fig. 5. Top of the list of immunologic factors that co-occur in sentences with the *Thalidomide*

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. 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 potential 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.

7 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 potentially contradicting information.

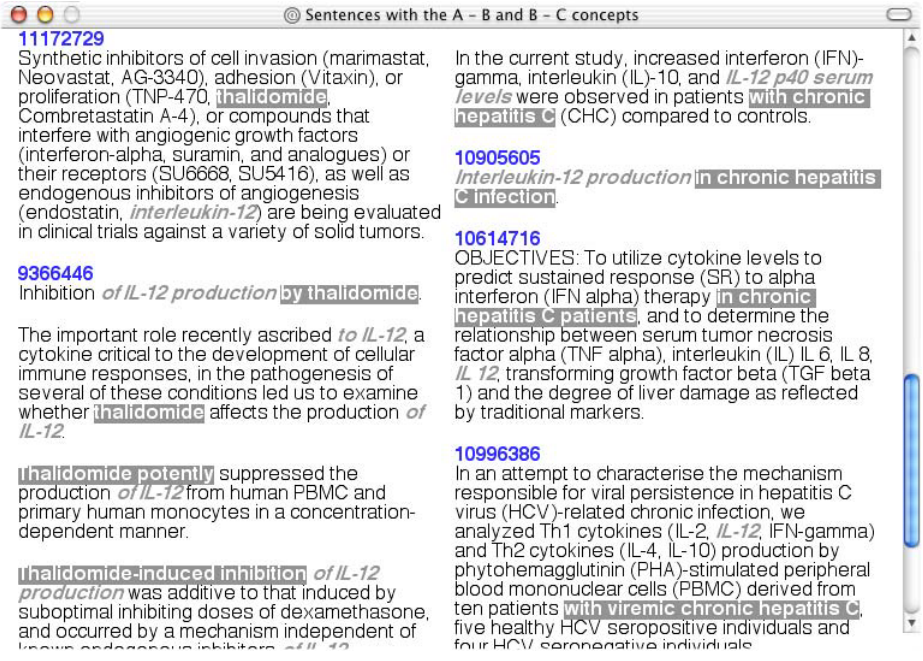

Fig. 6. Bibliographic information that suggests that chronic hepatitis C may benefit from thalidomide through IL-12 inhibition. The left column shows sentences in which A (thalidomide) and B(interleukin 12) concepts co-occur, the right column shows the relevant sentences for B and C(chronic hepatitis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

For instance, there is one MEDLINE citation that co-mentions thalidomide and myasthenia gravis and it claims that thalidomide is not effective in Lewis rats with 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, as this is the 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 be interesting to test. For instance, since thalidomide has some severe side effects, a clinical application may only be interesting in severe diseases or diseases for which there is no treatment at all. We concur with Smalheiser (2002) who views a literature-based discovery approach not as a replacement but as an added value to current hypothesis driven experimental research. Smalheiser envisions a

research environment in which informatics tools make hypothesis-driven research more efficient and productive.

There is some scepticism towards literature-based discovery and its potential for scientific research. Results are considered too obvious and once a 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..." originally intended to downplay the discovery, but actually resulting in yet another plausible hypothesis. This can be seen as a kind of activation of dormant knowledge in the mind of a scientist.

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. A paper describing the new potential uses of thalidomide resulting from literature-based discovery is currently under review for a biomedical journal.

Second, no one has denied the premise of the model, i.e., 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 (Spasser, 1997) 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 (Cory, 1997).

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. Even wider gaps between disciplines may result in interesting new insights.

Research in literature-based discovery has been acknowledged as important 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 (Spasser, 1997). Recently, however, a substantial National Institutes of Health 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, see also `http://arrowsmith.psych.uic.edu`). The goal of this project is to use informatics tools to optimize communication between neuroscientists and to connect individual research projects, data, and results. Researchers in five neuroscience laboratories will use a further developed version of ARROWSMITH to generate new hypotheses that they will test experimentally. 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, all at the University of Groningen, 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 discussions and 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

Aronson, A.R.: The effect of textual variation on concept based information retrieval. In: Proceedings of the Annual Symposium of American Medical Informatics Association AMIA-96, Nashville, TN, pp. 373–377 (1996)

Aronson, A.R.: Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap program. In: Proceedings of the Annual Symposium of American Medical Informatics Association AMIA-01, Washington, DC, pp. 17–21 (2001)

Cory, K.A.: Discovering hidden analogies in an online humanities database. Computers and the Humanities 31, 1–12 (1997)

Gordon, M.D., Lindsay, R.K.: 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, 116–128 (1996)

Hristovski, D., Stare, J., Peterlin, B., Dzeroski, S.: Supporting discovery in medicine by association rule mining in medline and umls. In: Proceedings of the Tenth World Congress on Medical Informatics, London, UK, pp. 1344–1348 (2001)

Langley, P.: The computational support of scientific discovery. International Journal of Human-Computer Studies 53, 393–410 (2000)

Lindberg, D.A.B., Humphreys, B.L., McCray, A.T.: The unified medical language system. Methods of Information in Medicine 32, 281–291 (1993)

Lindsay, R.K., Gordon, M.D.: Literature-based discovery by lexical statistics. Journal of the American Society for Information Science 50, 574–587 (1999)

Rikken, F.: Adverse drug reactions in a different context. Doctoral dissertation, University of Groningen, The Netherlands (1998)

Rikken, F., Vos, R.: How adverse drug reactions can play a role in innovative drug research. Pharmacy World & Science 17, 195–200 (1995)

Rindflesch, T.C., Aronson, A.R.: Ambiguity resolution while mapping free text to the UMLS Metathesaurus. In: Proceedings of the Annual Symposium of American Medical Informatics Association AMIA-94, San Francisco, CA, pp. 240–244 (1994)

Simon, H.A., Valdés-Pérez, R.E., Sleeman, D.H.: Scientific discovery and simplicity of method. Artificial Intelligence 91, 177–181 (1997)

Smalheiser, N.R.: Informatics and hypothesis-driven research. EMBO Reports 3, 702 (2002)

Smalheiser, N.R., Swanson, D.R.: Using ARROWSMITH: A computer-assisted approach to formulating and assessing scientific hypotheses. Computer Methods and Programs in Biomedicine 57, 149–153 (1998)

Spasser, M.A.: The enacted fate of undiscovered public knowledge. Journal of the American Society for Information Science 48, 707–717 (1997)

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

Swanson, D.R.: Migraine and magnesium: Eleven neglected connections. Perspectives in Biology and Medicine 31, 526–557 (1988)

Swanson, D.R., Smalheiser, N.R.: An interactive system for finding complementary literatures: A stimulus to scientific discovery. Artificial Intelligence 91, 183–203 (1997)

Valdés-Pérez, R.E.: Principles of human computer collaboration for knowledge discovery in science. Artificial Intelligence 107, 335–346 (1999)

Vos, R.: Drugs looking for diseases. Kluwer Academic Publishers, The Netherlands, Dordrecht (1991)

Weeber, M.: Literature-based discovery in biomedicine. Doctoral dissertation, University of Groningen, The Netherlands (2001)

Weeber, M., Klein, H., Aronson, A.R., Mork, J.G., de Jong-van den Berg, L.T.W., Vos, R.: Text-based discovery in biomedicine: The architecture of the DAD-system. In: Proceedings of the Annual Symposium of American Medical Informatics Association AMIA-00, Los Angeles, CA, pp. 903–907 (2000a)

Weeber, M., Vos, R., Baayen, R.H.: Extracting the lowest-frequency words: Pitfalls and possibilities. Computational Linguistics 26, 301–317 (2000b)

Weeber, M., Vos, R., Klein, H., de Jong-van den Berg, L.T.W.: 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, 548–557 (2001)

Weeber, M., Vos, R., Klein, H., de Jong-van den Berg, L.T.W., Molema, G.: Generating hypotheses by discovering implicit associations in the literature. a case report of a search for new potential therapeutic uses for thalidomide. Journal of the American Medical Informatics Association 10, 254–262 (2003)