Text-Based Discovery in Biomedicine: The Architecture of the *DAD*-system

Marc Weeber, MA*, Henny Klein, PhD*, Alan R. Aronson, PhD†, James G. Mork, MSc†, Lolkje T.W. de Jong - van den Berg, PhD*, Rein Vos, MD, PhD*‡

http://www.farm.rug.nl/dad/

* Social Pharmacy and Pharmacoepidemiology, Groningen University Institute for Drug Exploration, Groningen, The Netherlands.
† Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, MD.
‡ Health Ethics and Philosophy, Faculty of Health Sciences, University of Maastricht, The Netherlands.

Current scientific research takes place in highly specialized contexts with poor communication between disciplines as a likely consequence. Knowledge from one discipline may be useful for the other without researchers knowing it. As scientific publications are a condensation of this knowledge, literature-based discovery tools may help the individual scientist to explore new useful domains. We report on the development of the DAD-*system, a concept-based Natural Language Processing system for PubMed citations that provides the biomedical researcher such a tool. We describe the general architecture and illustrate its operation by a simulation of a well-known text-based discovery: The favorable effects of fish oil on patients suffering from Raynaud's disease [1].*

INTRODUCTION

Scientific knowledge grows at an enormous rate. Nowadays scientists are highly specialized researchers with exhaustive expertise on only a limited number of subjects. In the recent years, we have seen an upsurge in the (electronic) availability of scientific data, information, and knowledge, opening new areas for researchers to explore. Ironically, this availability has only increased the workload for the individual researcher: How does he find the desired information?

In the case of biomedicine, we observe that many data and information sources have become publicly accessible, with the genetic databases as the most extensive ones. Also, digested information, i.e. scientific knowledge, is abundant: Many thousands of scientific journals, of which many hundreds are available electronically, are potential sources for a researcher to browse. Fortunately, there are literature databases such as MEDLINE, with public access through interfaces such as PubMed [2], that comprise condensed information of these sources. However, the amount of the citations available in MEDLINE, 10,000,000, shows that it is humanly impossible to acquire all this knowledge.

For normal information needs, different query interfaces have been developed to MEDLINE. Using text words, Medical Subject Headings (MeSH), and many other characteristics, the user can find citations in MEDLINE that match his needs. Suppose that we are trying to find new leads to treat Raynaud's disease. We start by looking for reviews on this disease and retrieve 385 citations (as of February 2000). We will have to find typical phenomena, body processes, and characteristics that can be a site of action for a new therapy. Platelet aggregation is one such characteristic (the query `Raynaud's disease AND platelet aggregation`[1] results in 65 hits). But how can platelet aggregation be manipulated in order to have positive effects on Raynaud's disease? The PubMed query `platelet aggregation` returns 29,681 hits, of which many will be about substances that inhibit platelet aggregability. Acetylsalicylic acid (aspirin), for instance, is a thoroughly researched drug in this context (adding `acetylsalicylic acid` to the latter query results in 3,745 hits) and has also found its way as a potential treatment for Raynaud's disease. Less well-known platelet aggregation inhibitors, e.g., fish oil (`platelet aggregation AND fish oil` matches only 353 citations), may not have penetrated the Raynaud medical research community.

BACKGROUND

Professor Don R. Swanson (University of Chicago) modeled the latter example as follows: Between knowledge on a disease *C* and a therapeutic substance *A*, there is the link *B*, typically a physiological process. In different medical disciplines, knowledge on the relation *AB* and *BC* may be available, but the implicit connection *AC* may not be known yet. Swanson repeatedly showed that disconnected bodies of biomedical knowledge can be connected by studying their re-

[1] The `courier` font in this paper indicates PubMed query terms. Concepts are in *italics*, and semantic types are in **bold font**.

1067—5027/00/$5.00

spective literatures via this model [1,3,4]. He made his early discoveries by studying the literature intensively. Coincidentally, he formed the hypothesis that fish oil (*A*) may have a beneficial effect on Raynaud's disease (*C*). The literature showed him three general pathways (*B*) between fish oil and Raynaud's disease: Blood viscosity, platelet function, and vascular reactivity.

Together with Smalheiser, he proposed several other literature-based hypotheses that have been published in medical journals [5, 6]. Gordon and Lindsay followed Swanson in the literature-based discovery research by applying Information Retrieval techniques to Swanson's early discoveries [7, 8].

THE *DAD*-SYSTEM: A DISCOVERY TOOL

Our interest in text-based scientific discovery has led us to the development of the *DAD*-system, a Natural Language Processing (NLP) system that guides us in a two-step discovery process. Because we envision text-based discovery as a human-centered activity, our goal has been to codify a practical tool that assists the biomedical researcher in formulating and initially testing hypotheses. This implies that text processing and database connectivity should be of no concern to the user; however, in case of selection options, the tool should prompt the user with the relevant questions. Before unfolding the architecture of the *DAD*-system, we will briefly discuss the main issues in discovery tools.

Generating and Testing Hypotheses. In our *DAD*-system, we start the discovery process from a starting point *C*, a disease, for instance. We try to find interesting site of actions (*B*) in order to find a new lead *A*. In this phase, we have *generated* a new hypothesis. By examining the literature of both *A* and *C*, we *test* (strengthen or reject) this hypothesis. This two-step approach is similar to Gordon and Lindsay [7, 8], but different from Swanson whose discovery tool ARROWSMITH [9] tries to find a link *B* between known *A* and *C*-literatures (hypothesis testing).

Concepts instead of Words. We have decided to proceed beyond the actual text words: The units of analysis are UMLS Metathesaurus concepts [10]. The reason for this approach is three-fold. First, we are interested only in biomedically interesting concepts. We therefore do not need an ad hoc stop list of words with limited semantic content (e.g., determiners, prepositions, adverbs) and of irrelevant meaning (i.e., to biomedicine). Second, we would like to identify medical compound phrases: *Blood Pressure*, for instance, is a compound concept. The final reason to use UMLS concepts is that they have been assigned one or more semantic types. This allows the implementation of a semantic filter which is crucial for not getting lost in the abundance of possible pathways.

The Role of the User. Literature-based discovery is not an autonomous process. The discovery question is user-generated: On what subject does the user want to obtain new knowledge? Additionally, the filtering and selection of interesting *B* or *C*-concepts is user-dependent: Interesting in this case means interesting *according to the current knowledge and goals of the user*; it is the user who will have to make an interpretation of the computer-suggested list of possible pathways.

ARCHITECTURE

We have opted for a client-server model in which the client is any standard web browser. Figure 1 shows the architecture of the *DAD*-system; the oval box depicts the actual *DAD*-system. The resources the system uses, both the databases and the NLP tools, are outside the box. The dashed line in this figure represents the iterations through the system, which will be discussed in a subsequent section. The smiling faces represent the user and the bold arrows his interaction with the system. The current section describes the different resources and their connectivity.

Resources

PubMed. The *DAD*-system is a literature-based discovery system: Each discovery starts and ends with the literature. In the current implementation we use PubMed [2] as our main data source because of its wide coverage of biomedical sciences and its public availability.

MetaMap. All raw text, be it the user's query or the PubMed citations, must be translated, or mapped, to UMLS Metathesaurus concepts. MetaMap, a text-to-concept mapping program developed at the National Library of Medicine, has proven to be successful in NLP applications [11–13]. For the raw text analysis, 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 text.

UMLS Knowledge Sources. As the *DAD*-system's processing is concept-based, the UMLS Knowledge Sources [10] are interwoven with our tool. First, MetaMap is UMLS-based (thesaurus, lexicon), but also the query generation process discussed in the next section, uses the Knowledge Sources (KS) directly in

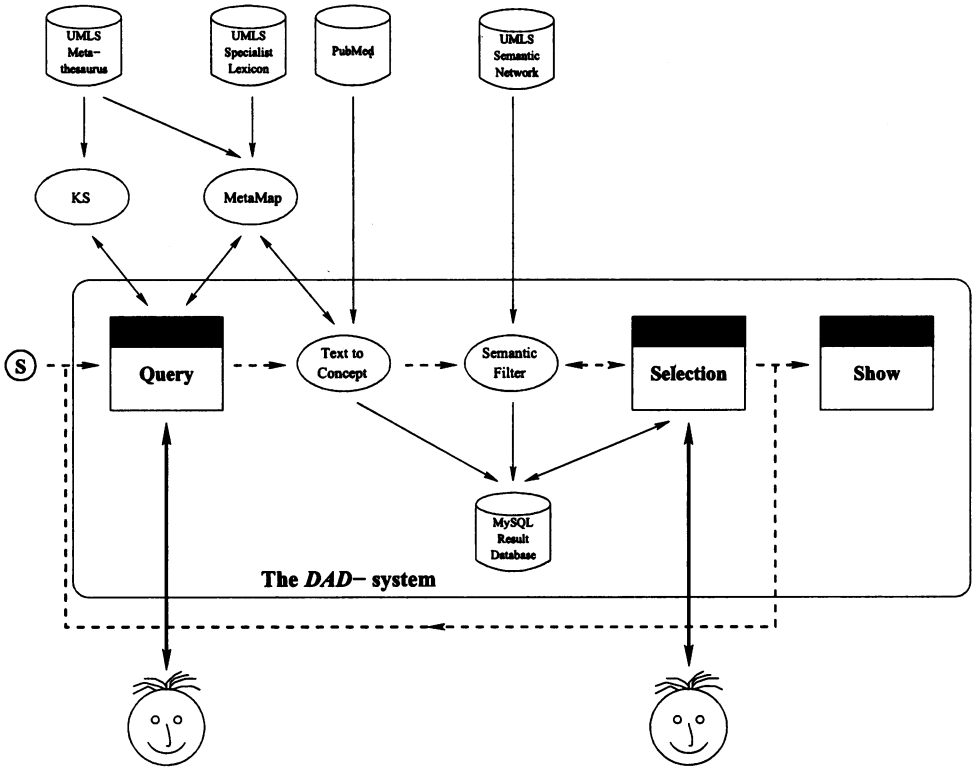

Figure 1: The *DAD*-system. The oval box represents the system itself, everything outside the box represent the resources. The dashed arrowed line indicate the trajectory through the system. See the Architecture section for a full description.

the synonym generation phase. The semantic information in the form of the 143 different semantic types plays the pivotal role in the filtering process.

SIMULATING RAYNAUD–FISH OIL

This section provides a simulation of Swanson's Raynaud's disease–fish oil discovery. Similar to Swanson, we use November 1985 as the upper publication date [1]. The example starts with the generation of a hypothesis. Our goal is to find new dietary factors that may affect Raynaud's disease. Subsequently, we will test the generated hypothesis. In Figure 1, we are at the starting point S.

Generating: $C \rightarrow B$

We are at the *DAD*-system's first human-computer interaction, the "Query" icon, where we enter `Raynaud's disease`. MetaMap maps this to the concept *Raynaud's disease*. We retrieve the synonym concepts *Raynaud's disease /phenomenon* and *Raynaud's syndrome* and generate the lexical variants for these concepts in a user-controled fashion. As we have defined a relation between two concepts as their co-occurence in a sentence, we need to backtranslate the concepts to a raw text representation that has to appear in the title or abstract of a MEDLINE citation. In the current example, we finish the query process with `raynaud` and `raynauds` as final PubMed query terms.

Using these final query terms, the *DAD*-system downloads the 1,246 relevant PubMed citations (the *C*-literature), communicates with MetaMap, and extracts and stores the relevant citation information in a local database. Subsequently, sentences that contain *Raynaud's disease* are selected and all concepts that appear in these sentences are put into a temporary database table. There are 1278 unique concepts in this table. Because we are interested in the physiological and functional aspects of Raynaud in this stage of discovery, we select the concepts that have the semantic types **Body Location or Region, Biologic Function, Cell Function, Phenomenon or Process, Physiologic Function,** or **Tissue**. This results in a list of 57 *B*-

concepts. Many of these concepts are related to blood factors; e.g., *Blood, Erythrocyte Deformability, Blood Viscosity, Platelet Adhesiveness*, and *Hemorheology*.

General concepts are likely to have many matching PubMed citations. The query `blood`, for instance, results in 800,000 PubMed citations up to 1985. This is clearly beyond the reach of our system. And the number of potential *C*-concepts in this huge collection is staggering so that we restrict our query to a limited number of specific concepts. Studying the list, expert users may observe (or already know) that blood viscosity and blood coagulation are related to Raynaud. We therefore select the concepts *Plasma viscosity level, Blood Viscosity, Platelet Adhesiveness, Platelet Aggregation*, and *Effects, Blood Coagulation*. The "Selection" icon in Figure 1 depicts this selection process. In case of doubt, the user can request for the context of a specific concept: The original sentences ("Show" icon). We follow the dashed line in the figure down and to the left, taking the selected *B*-concepts to a new query process.

Generating: $B \rightarrow A$

We are again at the left part of Figure 1: The query generation phase for the blood-related concepts. Directed by the *DAD*-system, we generate 25 query terms, e.g., `blood coagulation`, `blood viscosity`, `plasma viscosity`, `platelet adhesiveness`, and `platelet aggregation`. The *DAD*-system downloads and processes the 10,611 matching citations (*B*-literature). Sentences in which the selected *B*-concepts occur are put into a temporary table. There are 7,702 unique concepts in these sentences, of which 6,747 do not occur in the *C*-sentences. The latter concepts are the *A*-concepts. At this stage, we are interested in the dietary factors among these concepts. We therefore single out the concepts that have a semantic type of **Vitamin, Lipid, or Element, Ion, or Isotope**. The list of potential dietay factors consists of 206 concepts. There are many concepts that are lipid related, e.g., *Lipids, Triglycerides, Lipoproteins, Fatty Acids <1>* and *Dietary Fats*. We observe that five concepts related to fish oil and its active ingredients, *eicosapentaenoic acid, Fish Oil, Fatty Acids, Omega-3, maxepa*, and *omega-3 polyunsaturated fatty acid*, are higly ranked. Ranking is based on the number of pathways (*B*) between *C* and *A*.

By now, we can formulate the hypothesis that fish oil and its ingredients may have a positive effect on patients suffering from Raynaud's disease. Looking for additional fish oil concepts in the list, we find *Cod Liver Oil* and *salmon oil*.

Testing: $A \rightarrow B \leftarrow C$

At this stage, we try to strengthen (or reject) the generated hypothesis: Through which mechanisms or pathways can fish oil treat Raynaud's disease? We already have collected the literature on Raynaud (*C*-literature). Directed by the *DAD*-system, we generate PubMed queries for the above selected *A*-concepts. The *DAD*-system downloads the matching 463 citations and handles the processing. The sentences of the *A*-literature in which the selected *A*-concepts occur are put in a temporary database table. This table includes 1,795 unique concepts of which 479 are also in the *C*-literature temporary table. These 479 potential *B*-concepts are submitted to the same physiological/functional semantic filter as in the generating, $C \rightarrow B$ phase. This results in a list of 45 *B*-concepts. We find the already known concepts, but additionally, we find *vasodilation <1>, Veins, Capillaries*, and the prostaglandin *Dinoprostone*, which refers to a general "vascular reactivity" pathway found by Swanson [1]. We also observe additonal viscosity and platelet concepts: *Fibrinolysis, deformability*, and *rheology*.

The *DAD*-system provides an option to study these *B*-concepts in their *AB* and *BC*-context; the "Show" icon in Figure 1. The relevant sentences are presented as a juxtaposition, i.e., the system displays for each *B*-concept all *AB*-sentences next to the *BC*-sentences so that the user can assess the validity of the *AC*-hypothesis.

NEW APPLICATIONS

The previous section showed that our prototype of the *DAD*-system simulates Swanson's first discovery successfully. Parallel to re-discovering Swanson's cases for validation,[2] we pursue the following novel research targets.

Adverse Drug Reactions as Possible Pathways

Drugs are developed with one goal in mind: The treatment of a specific disease. Like any chemical compound, a drug will have a range of biological and physiological effects with the intended one as the only effect that has been researched extensively. However, many other effects, often called side-effects or Adverse Drug Reactions (ADRs), merit a closer examination for potential beneficial effects for other diseases [14]. A gastrointestinal drug with hypotensive effects, for instance, may be considered for use in hypertensive patients. Thus, we want to broaden the discovery process by including drugs as a target [15]. Because of our

[2]Recent experimentation showed that we were able to replicate Swanson's second discovery, the indirect influence of magnesium on migraine headaches. We are preparing a paper on this which also includes a strength and weakness analysis.

interest in ADRs, we have named our system the *drug-ADR-disease* or *disease-ADR-drug*-system, the *DAD*-system for short.

We are investigating the retrospective case of finasteride. This drug, originally intended for the treatment of benign prostatic hyperplasia, showed the ADR of hair growth. Recently, finasteride has been approved for use in patients with alopeica and male pattern baldness.

Epidemiology

We think the *DAD*-system to be a valuable addition to (pharmaco)epidemiology. The central issue in epidemiology is the association of a disease with risk factors. The actual mechanisms or pathways of these associatons are often not known. The *DAD*-system may help the epidemiologist in locating new risk factors and evaluating possible pathways between a risk factor and a disease. Our first steps in this research is the association between estrogen therapy and microalbuminuria. Both [16] and our own pharmacy databases show that patients using estrogens have a significantly higher risk to suffer from microalbuminuria. The pathways are not clear yet. The *DAD*-system is likely to be an effective tool to elicit the observed association.

Future Perspectives

The *DAD*-system uses PubMed as its source of scientific knowledge. For more specific discovery routes, however, better sources are available. We predominantly think of genetic databases that can be a source of knowledge to find genetic pathways of diseases. Finally, we envision text-based discovery as an extention to a standard MEDLINE literature search in order to reach beyond the scope and knowledge of the individual researcher.

CONCLUSION

We have implemented a text-based discovery system that can both generate and test novel hypotheses. The system can assess dietary-based hypotheses, such as Swanson's findings, and drugs-based hypotheses. However, the user can focus his search to any target by opting for a different semantic filter, which makes the *DAD*-system a versatile discovery tool for the biomedical expert.

REFERENCES

[1] Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect Biol Med 1986;30 (1):7–18.

[2] NLM. PubMed, 2000. `http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed`.

[3] Swanson DR. Migraine and magnesium: Eleven neglected connections. Perspect Biol Med 1988; 31 (4):526–557.

[4] Swanson DR. Somatomedin C and arginine: Implicit connections between mutually isolated literatures. Perspect Biol Med 1990;33 (2):157–186.

[5] Smalheiser NR, Swanson DR. Indomethacin and Alzheimer's disease. Neurology 1996;46:583.

[6] Smalheiser NR, Swanson DR. Calcium-independent phospholipase A2 and Schizophrenia. Arch Gen Psychiatry 1998;55:752–753.

[7] Gordon MD, Lindsay RK. 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 Am Soc Inf Sci 1996; 47 (2):116–128.

[8] Lindsay RK, Gordon MD. Literature-based discovery by lexical statistics. J Am Soc Inf Sci 1999;50 (7):574–587.

[9] Swanson DR, Smalheiser NR. An interactive system for finding complementary literatures: A stimulus to scientific discovery. Artif Intell 1997;91:183–203. `http://kiwi.uchicago.edu`.

[10] NLM. Unified medical language system knowledge sources, 2000. `http://umlsks.nlm.nih.gov/`.

[11] Rindflesch TC, Aronson AR. Ambiguity resolution while mapping free text to the UMLS Metathesaurus. In: *Proc Annu Symp Comput Appl Med Care 1994*. Philadelphia, PA: Hanley and Belfus, 1994; pp. 240–244.

[12] Aronson AR. The effect of textual variation on concept based information retrieval. In: *Proc AMIA Symp 1996*. Philadelphia, PA: Hanley and Belfus, 1996; pp. 373–377.

[13] Aronson AR, Rindflesch TC. Query expansion using the UMLS Metathesaurus. In: *Proc AMIA Symp 1997*. Philadelphia, PA: Hanley and Belfus, 1997; pp. 485–489.

[14] Rikken F, Vos R. How adverse drug reactions can play a role in innovative drug research. Pharm World Sci 1995;17 (6):195–200.

[15] Vos R. *Drugs looking for diseases*. Dordrecht, The Netherlands: Kluwer Academic Publishers, 1991.

[16] Ribstein J, Halimi JM, du Cailar G, Mimran A. Renal characteristics and effect of angiotensin suppression in oral contraceptive users. Hypertension 1999;33 (1):90–95.