Journal of Biomedical Informatics 2013;46(2):238-251.

A Graph-Based Recovery and Decomposition of Swanson's Hypothesis using Semantic Predications

Delroy Cameron[a,*], Olivier Bodenreider[c], Hima Yalamanchili[b], Tu Danh[b], Sreeram Vallabhaneni[b], Krishnaprasad Thirunarayan[a], Amit P. Sheth[a], Thomas C. Rindflesch[c]

[a]Ohio Center of Excellence in Knowledge-enabled Computing (Kno.e.sis)
[b]Biomedical Sciences Division
Wright State University, Dayton OH 45435, USA
[c]Lister Hill National Center for Biomedical Communications, National Library of Medicine
8600 Rockville Pike, Bethesda MD 20894, USA

Abstract

Objectives: This paper presents a methodology for recovering and decomposing Swanson's *Raynaud Syndrome–Fish Oil* Hypothesis semi-automatically. The methodology leverages the semantics of assertions extracted from biomedical literature *(called semantic predications)* along with structured background knowledge and graph-based algorithms to semi-automatically capture the informative associations originally discovered manually by Swanson. Demonstrating that Swanson's manually intensive techniques can be undertaken semi-automatically, paves the way for fully automatic semantics-based hypothesis generation from scientific literature.
Methods: Semantic predications obtained from biomedical literature allow the construction of labeled directed graphs which contain various associations among concepts from the literature. By aggregating such associations into informative subgraphs, some of the relevant details originally articulated by Swanson has been uncovered. However, by leveraging background knowledge to bridge important knowledge gaps in the literature, a methodology for semi-automatically capturing the detailed associations originally explicated in natural language by Swanson has been developed.
Results: Our methodology not only recovered the 3 associations commonly recognized as Swanson's Hypothesis, but also decomposed them into an additional 16 detailed associations, formulated as chains of semantic predications. Altogether, 14 out of the 19 associations that can be attributed to Swanson were retrieved using our approach. To the best of our knowledge, such an in-depth recovery and decomposition of Swanson's Hypothesis has never been attempted.
Conclusion: In this work therefore, we presented a methodology for semi-automatically recovering and decomposing Swanson's *RS-DFO* Hypothesis using semantic representations and graph algorithms. Our methodology provides new insights into potential prerequisites for semantics-driven Literature-Based Discovery (LBD). These suggest that three critical aspects of LBD include: 1) the need for more expressive representations beyond Swanson's *ABC model*; 2) an ability to accurately extract semantic information from text; and 3) the semantic integration of scientific literature with structured background knowledge.

Keywords: Literature-based Discovery (LBD), Swanson's Hypothesis, Semantic Predications, Semantic Associations, Subgraph Creation, Background Knowledge

1. Introduction

Literature-Based Discovery (LBD) is characterized by uncovering hidden but novel information implicit in noninteracting literatures. The field was pioneered by Don R. Swanson in 1986, through the well-known *Raynaud Syndrome-Fish Oil* Hypothesis *(RS-DFO)* [1]. Swanson serendipitously observed that dietary fish oils *(DFO)* appear to lower blood viscosity, reduce platelet aggregation and inhibit vascular reactivity (specifically vasoconstriction). Concomitantly, a reduction in blood viscosity and platelet aggregation, and the inhibition of vascular reactivity appeared to prevent Raynaud Syndrome *(RS)*; a circulatory disorder that causes periods of severely restricted blood flow to the fingers and toes [2]. Swanson therefore postulated that *"dietary fish oil might ameliorate or prevent Raynaud's syndrome."*

Remarkably, explicit associations between DFO and these intermediate concepts (i.e., blood viscosity, platelet aggregation and vascular reactivity) had long existed in the literature. Likewise, explicit associations between the intermediate concepts and RS had been well documented. The serendipity in Swanson's Hypothesis lies in the fact that no explicit associations linking DFO and RS had been previously articulated. To arrive at this discovery, in November 1985, Swanson obtained an initial seed set of 489 articles and performed a Dialog® Scisearch using Raynaud and Fish Oil terms, on titles and abstracts of MEDLINE and Embase (Excepta Medica) citations. He found that among the 489 articles, only four articles contained cross-references spanning both the DFO and RS group. Among these four, he found that only two [3, 4] associated RS

[*] Corresponding Author. Tel.: +1 937 775 5213; fax: +1 937 775 5133
Email address: delroy@knoesis.org (Delroy Cameron)

with DFO. Swanson speculated that this phenomenon of *noninteracting literatures*, alludes to the existence of *undiscovered public knowledge* [5]. He exploited his awareness of the existence of such undiscovered *semantic associations* [6] across noninteracting literatures and investigated several other scenarios [7, 8, 9, 10, 11] (along with Smalheiser) that later led to new scientific discoveries.

Swanson grounded his observations in a paradigm refered to as the *ABC model* [1]. This model states that new knowledge can be discovered between two concepts (A,C) from noninteracting literatures, if hidden associations involving some intermediate concept (B) can be uncovered This seminal model has revolutionized the field of LBD, and has been used both to recover many of Swanson's original hypotheses [12, 13, 14, 15, 16, 17, 18] as well as to propose new hypotheses [19, 17, 20, 21].

However, while Swanson did indeed provide detailed explanations for his observations, he presented no discussion on a framework for finding such details, automatically or otherwise. Instead, Swanson's ABC model has been used extensively to arrive at high-level conclusions. Hence, much of the early LBD research [22] used IR techniques [13, 12, 21, 23, 24, 25] to illustrate the effectiveness of the ABC model for LBD. The Arrowsmith search tool [23, 24], developed by Smalheiser and Swanson, epitomizes LBD achieved using this ABC-IR framework. The underlying philosophy behind ABC-IR has been that, in scenarios in which both source (A) and target (C) are known (*closed discovery*), new discoveries will arise from intermediates that 'frequently co-occur' with source and target. Hence, techniques such as surface form normalization [25, 14], text-to-concept mapping [16, 14, 25], term counts and relative frequencies [13, 12, 25] have been used extensively for finding and ranking intermediates. The main issue with term co-occurrence approaches however, (whether strictly lexical or concept-based) is that while they often succeed in finding intermediate (B) concepts, they provide no insight into the nature of the relationships among the concepts. For example, while the terms "DFO," "platelet aggregation" and "RS" may frequently co-occur in some corpus, their co-occurrence does not explicitly reflect the association that *(DFO INHIBITS platelet aggregation)* and an increase in *(platelet aggregation CAUSES RS)*.

This limitation has far reaching implications in the biomedical domain. Smalheiser [26] recently noted that next generation LBD requires more expressive representations beyond the ABC model. Ahlers et. al. [19] makes the specific observation that in treatment of diseases for example, *"Drug therapies are often used effectively, even though the exact cause of action may be either poorly understood or unknown."* Biomedical researchers are therefore not only interested in mere co-occurrence relationships, but also in understanding the mechanisms of interaction and causality relationships among concepts. In the previous example (p_1), the detailed association that p_1=*(DFO STIMULATES Epoprostenol)*→*(Epoprostenol ISA Prostaglandin)*→*(Prostaglandin INHIBITS platelet aggregation)*→*(platelet aggregation CAUSES RS)* is of more interest to researchers because it conveys causality.

IR techniques therefore may fail to provide context and domain semantics in such critical scenarios. Other approaches, including those based on techniques such as Latent Semantic Indexing [27] and concept-based link analysis [14, 28] also suffer the same drawback. Approaches based entirely on the ABC paradigm and the idea that one level of intermediates is sufficient for LBD, offer limited coverage across the relevant associations to the discourse. Naturally, relevant information may exist in longer chains of concepts semantically connected.

Semantics-based approaches to LBD [19, 17, 18, 29, 20] therefore aim to provide context as well as improve coverage. To achieve this, they rely on assertions *(or semantic predications)* extracted from the literature. Semantic predications are binary relations of the form *(subject, predicate, object)*, where the *predicate* expresses a relationship between the *subject* and the *object*. For example, in the semantic predication *(DFO INHIBITS platelet aggregation)*, the *predicate* "INHIBITS" expresses the relationship between the *subject* "DFO" and the *object* "platelet aggregation." Wilkowski et. al. [20] was among the first to demonstrate the inherent value in using semantic predications to move LBD beyond the canonical ABC model. Without loss of generality, Wilksowski proposed that the ABC model can be decomposed into a more granular model in which several intermediate concepts may be required to expound associations. We refer to Wilkowski's logical extension of Swanson's ABC model, as the *AnC model (pronounced ants)*, in which n=(B_1, B_2, . . . B_m).

While more expressive than the ABC model, the *AnC model* itself is not foolproof. Consider an extension of the previous example (p_1), consisting of two associations (p_2, p_3) instead of one. Suppose the first association states that p_2=*(DFO STIMULATES Epoprostenol)*→*(Epoprostenol TREATS RS)* and the second states that p_3=*(DFO CONVERTS_TO Prostaglandin (PGI$_3$))*→*(Prostaglandin (PGI$_3$) INHIBITS platelet aggregation)*→*(platelet aggregation CAUSES RS)*. Further, suppose that these two associations (p_2, p_3), along with the association from the previous example (p_1) and a host of other associations (p_4, p_5, . . . , p_i) are part of a labeled graph of semantic predications *(called a predications graph)*. These two associations (p_2, p_3), connected by their common vertices in this predications graph, will naturally form a *subgraph* (Figure 1, left). However, unlike example (p_1), in this scenario the role of Epoprostenol and platelet aggregation together, in treating RS has now been obscured. Instead, it requires background knowledge (whether from structured sources or from domain experts) to bridge the gap. It can therefore be established and expressed through another semantic predication that *(Prostaglandin (PGI$_3$) ISA Epoprostenol)*. Given this additional information from background knowledge, we can then conclude through transitivity that there exists at least one instance in which an *Epoprostenol*, namely *(Prostaglandin (PGI$_3$)* INHIBITS platelet aggregation), as the mechanism for treating RS (Figure 1, right). From this example, it is clear that while both the ABC model and the *AnC model* may be sufficient for LBD, the construction of relevant subgraphs that leverage background knowledge (as in Figure 1) play a critical role in supplementing and complementing both models. This example highlights the first and third points put forth in the *Conclusion* Section of the abstract, which state that the future of semantics-based LBD requires: 1) more *expressive representations beyond Swanson's ABC model* and also 3) *the semantic integration of scientific*

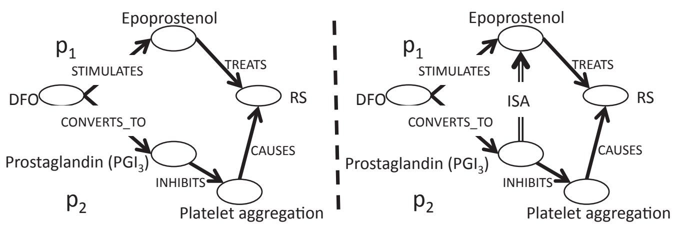

Figure 1: From Associations to Subgraph to Background Knowledge

literature and structured background knowledge.

In this work, we therefore investigate the use of semantic predications, semantic associations, expressive subgraphs and background knowledge for LBD, using graph-based techniques. We make the following specific contributions:

- We are the first to conduct a detailed decomposition of Swanson's *RS-DFO* Hypothesis into 19 associations, compared with recovery approaches limited to only the 3 well known associations. Hence, we obtain more coverage across Swanson's Hypothesis.

- Similar to Wilkowski, we extend the classical ABC model into an more expressive model which we term the *AnC model*, leveraged for semantic association generation. The expressive associations resulting from this extension provide detailed explanations needed to understand the causality relationships and mechanisms of interaction among concepts.

- Further, we illustrate the need for subgraphs instead of mere associations, especially in scenarios when associations alone are not sufficiently expressive to adequately expound connections among concepts.

- Finally, we show that when subgraphs themselves are not sufficiently expressive, semantically integrating background knowledge to bridge knowledge gaps in the scientific literature is crucial for LBD.

2. Related Work

Hristovski was among the first to present ideas advocating the use of semantic predications for LBD. His pattern-based approach [17, 30] leverages both semantics (from SemRep [31]) and term co-occurrence (from BioMedLEE [32]) to recover existing knowledge as well as to suggest new relationships between concepts. He uses *discovery patterns* specified *a priori*, to find potentially interesting associations. For example, if some disease (A) and some drug (C) have opposing effects on the same intermediate concept (B) then he speculates that drug (C) *Maybe_Treats* disease (A). In the case of the *RS-DFO* Hypothesis, since DFO “reduces” blood viscosity and platelet aggregation, and RS occurs when blood viscosity and platelet aggregation levels increase, Hristovski infers that DFO *Maybe_Treats* RS. The “reduces-increases” pattern is used to imply *Maybe_Treats*. The limitation of Hristovski's pattern-based approach is that it cannot be easily extended to accommodate complex patterns. In fact, since the patterns are restricted to the proverbial ABC paradigm, interesting and more complex associations will go undetected.

Ahlers et. al. [19] also applies semantic predications to LBD, leveraging Hristovski's idea of *discovery patterns*. Under the ABC paradigm they infer the *Maybe_Disrupts* predicate. The argument is that, an antipsychotic drug (A) *Maybe_Disrupts* cancer if the drug (A) *INHIBITS* some intermediate bioactive, pathological and/or pharmacologic concept (B) which *CAUSES, PREDISPOSES* or is *ASSOCIATED_WITH* cancer (C). The ‘inhibits-causes’ pattern is used to infer *Maybe_Disrupts*. Ahlers discovered five interesting intermediate concepts relating antipsychotic drugs and cancer using this pattern. In this work, we eliminate the need to know and specify discovery patterns *a priori* by implementing a graph traversal algorithm agnostic to specific patterns.

Aside from pattern-based approaches, some *discovery support systems* use various heuristics to provide focused search and browsing to support semantics-based LBD. Cohen et. al. [29] implemented Epiphanet, which visualizes associations of varying lengths between concepts based on chains of predications between them. Epiphanet uses the notion of random reflexive indexing (RRI) to establish connections between concepts that do not necessarily co-occur in text but are related in the corpus, based on normalizing lexical variants. We tested the “logical leap” feature of Epiphanet with DFO variants (i.e., the UMLS concepts Fish oil - dietary (FOD), Fish Oils (FO) and Eicosapentaenoic Acid (EPA)) as the source and Raynaud Phenomenon (RP) as the target. Unfortunately no associations were found that directly connected source and target.

In previous work, we [33] implemented a discovery support system, later enhanced by Kavuluru et. al. [34] and called Scooner, that also leverages structured knowledge to facilitate LBD. Users can perform assertion-driven (semantic) browsing by first selecting one or more annotated concepts/subjects from one MEDLINE article, then selecting a relevant predicate (p) for which (s) is the subject, and finally selecting an object (o) grounded in a different MEDLINE article. LBD may occur after users analyze associations or *semantic trails* created through such contextual navigation. We tested Scooner with Eicosapentaenoic Acid (EPA) as the source and Raynaud Disease (RD) as the target and were unable to recover Swanson's Hypothesis within a reasonable time frame. We acknowledge that while the potential for replicating Swanson's *RS-DFO* Hypothesis may exist using Scooner, our efforts to do so in this particular scenario were unsuccessful.

Pratt and Yetisgen-Yildiz [16, 28] implemented a system called LitLinker that uses UMLS concepts, UMLS semantic network types and the idea of *level of support* for open discovery. In *open discovery* only the starting concept and some intermediate are known, but not the target. By eliminating irrelevant intermediate concepts *(i.e., linking concepts)*, based on low level of support from the corpus, Pratt successfully retrieved several intermediate concepts from Swanson's *Magnesium-Migraine* Hypothesis [7]. She first measured the support for linking concepts by estimating its likelihood based on the number of titles containing it in the corpus, and then filter target concepts (under the ABC paradigm) based on the centrality score of their linking concepts. This approach may lack coverage, since ranking target concepts based on centrality scores may eliminate interesting but rarely mentioned intermediate concepts. Our graph-based approach that leverages semantics is more flexible and adaptable to ensure better coverage than Pratt's approach.

Wilkowski et. al. [20] also proposed graph-based ideas for open discovery using the notion of *discovery browsing* together with *degree centrality* using the predications graph. Discovery browsing seeks to present users with discovery patterns in a user-friendly way. This is achieved by ranking predications according to the degree centrality of their subject and object. Ideally, a user can traverse the predications graph by visiting high centrality vertices (predications) in succession. Similar to Pratt [16] however, Wilkowski's approach may eliminate interesting outliers. Note that Wilkowski was the first to materialize an extension of the ABC model in which the intermediate concept (B) was a set, instead of just a single concept.

In previous work, we [35] also describe a graph-theoretic application of the predications that has implications on Question Answering (QA) as well as LBD. By using the predications to connect documents that answer complex questions, they suggest that associations may inherently contain knowledge that could lead to new scientific hypotheses. They use a modified depth first search (MDFS) algorithm along with various heuristics to make logical leaps between concepts. Their notion of *knowledge abstraction* leverages semantic relatedness among concepts from background knowledge; a theme reiterated here.

Finally, Hristovski implemented a discovery support system called BITOLA [18] that also leverages graph theory and term co-occurrence statistics for LBD. BITOLA is applicable in both closed and open discovery scenarios. It uses *confidence* and evidence-based *support* to rank pairs of concepts from corpus statistics. We tested BITOLA for closed discovery using Eicosapentaenoic Acid (EPA) as the source, Raynaud Disease (RD) as the target, and with "Cell Function" as the *a priori* semantic type of the intermediate. We found the intermediates, "platelet aggregation" and "erythrocyte deformability" as the #1 and #2 ranked concepts respectively, among only 24 intermediates. This result is very impressive. We then tested BITOLA for open discovery using EPA as the source, "blood viscosity" as the intermediate and "Eicosanoid" as the semantic type for the target. We found the target Epoprostenol at position 3 out of 21, which is also remarkable. The problem is that "Epoprostenol" cannot be further explored without constructing another query. This restriction is due to the ABC model. As evidenced by the role of "Epoprostenol," "Prostaglandins" and "platelet aggregation" in treating RS (discussed in Section 1 Figure 1), provisions for exploring more detailed associations is essential. We believe this scenario corroborates the need for the *AnC model*, to better explicate associations between concepts.

3. Method

Our graph-based framework for recovering and decomposing Swanson's *RS-DFO* Hypothesis involves five tasks; 1) literature selection and preprocessing; 2) semantic predication extraction from the selected literature; 3) building the predications graph from the extracted semantic predications; 4) applying a search algorithm for semantic association generation; and 5) subgraph creation using relevant associations, and background knowledge where appropriate. We begin by discussing literature selection and preprocessing in the next Section.

3.1. Literature Selection and Preprocessing

The sixty-five articles cited in Swanson's original *RS-DFO* paper [1] were selected as the baseline dataset for our approach. Such a minimal dataset was selected mainly because we are interested in assessing the feasibility of using semantic predications, semantic associations, expressive subgraphs, background knowledge and graph algorithms for recovering an existing hypothesis.

We believe that if our techniques can be successfully applied to rediscover existing knowledge, then moving from rediscovery to actual knowledge discovering will require a recalibration of our approaches to deal mainly with scalability. Proving that existing knowledge can be recovered on a controlled dataset is an important initial step in devising techniques that will be effective on larger heterogeneous datasets. Hence, we created this minimum baseline of the 65 articles, by manually searching PubMed, Google Scholar, Mendeley, Wiley Online Library, Science Direct, etc. for abstracts and full text articles. Altogether, one article (citation#60) was not found at all, while another article was written in French (citation#64), although an English version consisting of its title and abstract was retrieved from PubMed. Twelve articles consisted of titles and abstracts only (citation#4,5,22,23,25,27,37,40,43,51,57,63), and six articles (citation#1,8,46,56,59) contained only titles and full text but no abstract. We subdivided this initial baseline into two sets. Baseline 1 (B1) consisted of titles, abstracts and full text of all articles while Baseline 2 (B2) consisted of only titles and abstracts of all articles. The choice of the two datasets was motivated by the desire to determine, to what extent titles and abstracts may be adequate for supporting LBD. Since MEDLINE provides only titles and abstracts for approximately 21 million scientific papers, this was an important undertaking.

Since many articles in the B1 dataset were images rather than text, we used an optical character recognition (OCR) text extractor called *The Tesseract OCR engine* [36] to convert PDF articles to text. We then converted non-ASCII characters to ASCII, since tesseract did not always accurately parse non-ASCII characters (such as α). A resulting subtask was the need to resolve end-of-line hyphenation and text wrapping. For instance, the word "polyunsaturated" could be expressed as "polyunsatur-$\backslash$*n*nated" if it appears at the end of a line in the PDF version of the article. Since it is necessary to distinguish between standard hyphenated words and those originating from PDF text wrapping, we obtained frequency counts for hyphenated words in the corpus. We then removed hyphens

Table 1: Dataset and Predications graphs

Dataset	#Semantic Predications	#Entities
B1:Full Text	4438	1078
B2:Title/Abstract	389	193

from words that were end-of-line words that fell below some empirically observed threshold frequency *(threshold=5)*. On the contrary, since the B2 dataset was collected from PubMed, neither the conversion to ASCII nor text wrapping issues surfaced while creating that dataset.

Due to space limitations, only some datasets and experimental results are listed in this manuscript. All datasets and experimental results are available online[1]. In the next Section we discuss how the semantic predications were extracted using these two baseline datasets B1 and B2.

3.2. Predication Extraction

Work on semantic predication extraction spans more than two decades of research by Thomas C. Rindflesch at NLM. The linguistics-based semantic predication extractor called *SemRep*, is now available as a web service exposed by NLM, through the Semantic Knowledge Representation (SKR) project. An abbreviated version of the SemRep output is given below, applied to title of a MEDLINE article *[PMID6130329]*:

```
Text — Intermittent epoprostenol (prostacyclin) infusion in patients with Raynaud's
syndrome.
entity — C0205267 — Intermittent — Intermittent
entity — C0033567 — Epoprostenol — epoprostenol
entity — C0033567 — Epoprostenol — prostacyclin
entity — C0574032 — Infusion procedures — infusion
entity — C0030705 — Patients — patients
entity — C0034734 — Raynaud Disease — Raynaud's syndrome
predication — Infusion procedures — TREATS — Patients
predication — Raynaud Disease — PROCESS_OF — Patients
predication — Infusion procedures — TREATS(INFER) — Raynaud Disease
```

Note that the labels “epoprostenol” and “prostacyclin” in the title, were both normalized to the concept “Epoprostenol (C0033567)” among the recognized entities in lines 3 and 4 of the output. We used the SKR Web API[2] to access SemRep to extract semantic predications from the baseline datasets. Extracted semantic predications were subsequently used to build the predications graphs.

3.2.1. Postprocessing

We made the following observation while processing the SemRep output. SemRep erroneously parses the phrase “vascular reactivity” (which maps to UMLS concept (vascular reactivity-C1660757)) into two tokens: 1) the first is “Vascular,” which maps to UMLS concept (Blood Vessels-C0005847) and 2) the second is “Reactivity,” which maps to UMLS concept (Reactive-C0205332). This problem was first detected by Hristovski et. al. in their attempts at rediscovering Swanson’s *RS-DFO* Hypothesis [17, 30]. Consequently, Hristovski only reported results on platelet aggregation and blood viscosity and ignored vascular reactivity altogether. We realized that there were only 26 unique mentions of vascular reactivity (and its lexical variants) in this corpus. Hence, we manually augmented the SKR output with semantic predications SemRep failed to extract. Such semantic predications are distinguished by appending the postfix “_MAN” to their UMLS predicates. For example, INHIBITS_MAN represents the manually identified UMLS predicate INHIBITS.

This action provided us the flexibility to first normalize the UMLS concept “vascular reactivity” to UMLS concept (Vascular constriction (function)-C0042396), since Swanson uses vascular reactivity in his manuscript [1] in reference to vasoconstriction. Further, since vasoconstriction and vasodilation convey opposite semantics, for each mention of vasoconstriction we added another semantic predication for vasodilation, containing the negating predicate, and vice versa. For example, since (Epoprostenol INHIBITS vasoconstriction) we included the semantic predication which states that (Epoprostenol CAUSES vasodilation) which has the negating predicate CAUSES for the predicate INHIBITS. From the manual annotations to vascular reactivity dataset (also available online), we obtained 76 semantic predications from B1 but only one from B2. Overall, our manual extraction highlights the second point put forth in the *Conclusion* Section of the abstract, which states that the future of semantics-based LBD warrants *an ability to accurately extract semantic information from text*.

In the next Section, we discuss the formalism and construction of the predications graphs using the SemRep extracted semantic predications and those manually added from the vascular reactivity dataset.

3.3. Predications Graph

We formally articulated the notion of a predications graph in [35], but we revisit it here for completeness. Using set notation, let $S(d_i)$ be the set of semantic predications associated with article d_i. If t denotes a semantic predication in d_i and D is the set of articles $\{d_1, d_2 \ldots, d_n\}$ then,

$$\text{For any } t = (s_t, p_t, o_t), \text{let } D(t) = \{d \mid t \in S(d)\} \tag{1}$$

be the set of corresponding articles that contain the semantic predication t and $S(D)$ be the set of all semantic predications associated with articles in D. That is,

$$S(D) = \bigcup_{i=0}^{|D|} S(d_i). \tag{2}$$

The semantic predications in $S(D)$ for a set of articles D naturally form a directed labeled graph, denoted $G_{S(D)}$, in which the subject and object of each semantic predication is a vertex,

[1] Datasets - `wiki.knoesis.org/index.php/Obvio#RS-DFO_Hypothesis`

[2] SKR Web API-`http://skr.nlm.nih.gov/SKR_API/`

and the predicate is a labeled edge from subject to object. This graph is called the *predications graph*.

Table 1 shows that the predications graph for B1 consisted of 4434 semantic predications and 1077 unique concepts, while the predications graph for B2 contains 388 semantic predications and 192 unique concepts. Fortunately, such small graphs impose no major demands for run-time optimizations to improve our search algorithm. In the next Section, we revisit the notion of a semantic association originally articulated by Anyanwu et. al. [6, 37], and discuss the algorithm for extracting such semantic associations from the predications graphs.

3.4. Semantic Association Generation

As established by the associations (p_1, p_2, p_3) in example 1, from Figure 1, Swanson's ABC model may be insufficient for LBD in cases when complex relationships exist among concepts. Expressive associations based on the *AnC model* are imperative if such complex associations will be captured. However, extracting expressive semantic associations that are relevant, interesting, plausible, and intelligible from an arbitrary labeled predications graph, is not trivial. Furthermore, the very notion of an association must be well understood if we are to capture such associations. We therefore revisit the notion of a semantic association here, and then discuss an approach for efficiently extracting them.

3.4.1. Semantic Association

Anyanwu et. al. [6, 37] defines a semantic association between two vertices (v_i, v_j) in terms of property sequences and joined property sequences in which the vertex (v_i) may be the origin of a property sequence, and the vertex (v_j) may also be the origin or terminus of another property sequence. Informally, a semantic association is a path connecting concepts through labeled edges in a directed graph, such that paths can be joined together on common vertices to give rise to more complex semantic associations. However, Anyanwu's definition does not make it explicit that a semantic association may contain sequences of edges in any direction. Consider for example, the association which states that p_4=*(Eicosapentaenoic Acid → INHIBITS → Vascular Constriction ← INHIBITS ← Nifedipine → TREATS → Raynaud Phenomenon)*. A plausible conclusion is that (Eicosapentaenoic Acid TREATS Raynaud Phenomenon) by inhibiting Vascular Constriction. We arrive at this conclusion using Hristovski's notion of a discovery pattern discussed in Section 2. Since Nidefipine has an inhibiting effect on Vascular Constriction (in the opposing direction), and also TREATS Raynaud Phenomenon, we can speculate that (Eicosapentaenoic Acid Maybe_Treats Vascular Constriction), since it too INHIBITS Vascular Constriction.

From this scenario, it is therefore clear that a semantic association must allow sequences of edges if we are to truly capture informative associations. Informally then, a semantic association is a path connecting concepts through labeled edges, such that paths can be joined together on common vertices to give rise to more complex semantic associations. Formally, a *semantic association* is defined as follows: Given a directed graph G=(V, E), where V=$\{v_0, v_1, \dots v_n\}$ is the vertex set and E=$\{e_0, e_1, \dots, e_m\}$ is the set of labeled edges, a semantic association $p_{(s,t)}$ exists between the vertex (s) and the vertex (t) if, for some arbitrary set of vertices V_p = $\{v_{p_0}, v_{p_1}, \dots, v_{p_d}\}$, $V_p \subset V$, for all vertex pairs (v_i, v_j) in V_p, j=i+1 or j=i-1, for all 1≤ i ≤ d.

Using this notion of a semantic association, we can then exploit the idea of *reachability* to generate semantic associations using the predications graphs.

3.4.2. Reachability

Reachability is the notion of being able to get from one vertex to some other vertex in a directed graph [35, 38]. Given our definition of a semantic association, a vertex t is reachable from another vertex s, if the two vertices are semantically associated. In general terms, the set of all semantic associations between all pairs of vertices (v_i, v_j) in the vertex set V of the graph G, is the transitive closure *(or reachability relation)* of the entire graph G. However, for our purposes we are only concerned with the transitive closure between the vertex pairs (s, t), where s=DFO and t=RS. Since there are various manifestations of DFO and RS in the UMLS, to compute this transitive closure we selected the following UMLS concepts: Fish Oil - dietary (FOD), whose CUI[3] Fish Oils (FO), whose CUI is C0016157 and Eicosapentaenoic Acid (EPA), whose CUI is C0000545 as synonyms for DFO. Similarly, we used the concept Raynaud Disease (RD), whose CUI is C0034734 and the concept Raynaud Phenomenon (RP), whose CUI is C0034735 as synonyms for RS. In the next Section, we formally define the reachability relation between two vertices.

3.4.3. Reachability Relation and Predication Selection

We obtained the transitive closure between a vertex pair using the classical Depth First Search (DFS) algorithm to traverse the predications graph according to our notion of a semantic association. We set variants of DFO as the root and variants of RS as the terminal. Hence, the input to the algorithm is various sets of (DFO, RS) vertex pairs, and the output is a set of semantic associations or the reachability relation between the vertex pairs. Informally, the reachability relation is the subgraph formed by the transitive closure between two vertices in a directed graph (ignoring directionality), which is the set of all semantic associations (by our definition) between the two vertices. Formally:

The Reachability Relation between the vertex pair (s, t), denoted R, is a subgraph R=$(V_r$,$E_r)$ where $R \subset G_{S(D)}$, $V_r \subset V$ and $E_r \subset E$, such that R is the transitive closure $P_{(s,t)}$=$\{p_1, p_2, \dots, p_k\}$ of associations for the vertex pair (s, t), where $p_k(V_r)$ is the vertex set of the k^{th} association in $P_{(s,t)}$, such that $p_k(V_r)$ = {s=$v_{k0}, v_{k1}, \dots, v_{k|p_k(V_r)|}$=$t$} for all (v_i, v_j) and j=i+1, or j=i-1, for all 1≤ i ≤ $|p_k(V_r)|$, and 1≤ k ≤$|P_{(s,t)}|$.

[3] Concept Unique Identifier or CUI - a identified by a unique identifier used to distinguish UMLS concepts is C0556145,

We initially selected a maximum depth of 3 as a stopping condition for the Depth First Search (DFS) algorithm, and observed a maximum running time of 1 second to generate the reachability relation. At depth 4, the maximum running time increased to 16 seconds, and at depth 5, the maximum was 4 minutes. Empirically, our experiments (discussed in Section 4) suggest that beyond the maximum depth of 3, the associations returned by the algorithm while many, produce diminishing returns. Hence, we report here on associations generated to the maximum depth of 3.

Further, to obtain these optimized running times, we excluded associations containing generic concepts, such as "Disease" and "Patient." The rationale is that such semantic predications are not very insightful. For example, the two semantic predications (Raynaud Disease ISA Disease) and (Raynaud Phenomenon PROCESS_OF Patients) are not very informative. Given that (Epoprostenol TREATS Diseases) and (Raynaud Disease ISA Disease) is certainly not sufficient evidence to reasonably conclude that (Epoprostenol TREATS Raynaud Disease). By the same argument, we also excluded associations containing semantic predications involving *weak predicates* (of little relevance in this context), such as PROCESS_OF, PROCESS_OF(SPEC), ADMINISTERED_TO, ADMINISTERED_TO(SPEC), PART_OF, ASSOCIATED_WITH, COEXISTS_WITH, TREATS(INFER). We added the TREATS(INFER) predicate to this set because, semantic predications containing the TREATS(INFER) predicate are inherently weaker that those containing the TREATS predicate. The stronger of these two is sufficient. In the next Section, we discuss the creation of relevant subgraphs using the reachability relations extracted from the predications graphs, using the DFS algorithm.

3.5. Subgraph Creation

We first manually identified a number of associations articulated by Swanson in his original *RS-DFO* paper [1]. To identify the associations, we created chains of semantic predications based on statements in the text. For example Swanson stated in the *RS-DFO* paper [1], Section: *The Effects of Dietary Fish Oil on Blood Viscosity, Platelet Function and Vascular Reactivity, page 4* that: *"It is known that EPA can suppress platelet aggregation by several different mechanisms, though which among them are of greatest importance is not known."* From this sentence we created a semantic predication which states that Dietary Fish Oil⟶INHIBITS⟶platelet aggregation, since EPA and Eicosapentaenoic acid are synonymous with dietary fish oil. The term "suppress" corresponds to the UMLS predicate "INHIBITS," and the term "platelet aggregation" is used as the object, since it is a known UMLS concept. By chaining together such predications, we identified three *primary associations* (shown in Table 3, ID: #1, #2, #3) involving platelet aggregation, blood viscosity and vascular reactivity, which are commonly discussed in biomedical literature and recognized as Swanson's *RS-DFO* Hypothesis. Additionally, we identified eight *supplementary associations* (Table 3, ID:1a on platelet aggregation, Table 3, ID:2a-c on blood viscosity and Table 3, ID:3a-d on vascular reactivity) which expound the primary associations in detail. Then, we identified eight *secondary associations* (Table 5, ID:1.1-2, 2.1-3, & 3.1-3) which provide associations between other concepts and RS (such as Ketanserin, Nifedipine and Alprostadil), but do not necessarily link DFO and RS directly.

From the collection of all reachability relations, we then manually constructed subgraphs by grouping semantic associations deemed relevant to each of Swanson's original claims. We then inspected each subgraph (denoted $G_{P_i^{DFS}}$) and used domain expertise to provide background knowledge when necessary. Since each subgraph from the DFS output either contains the same knowledge as Swanson's association (denoted p_i^S) or requires background knowledge for extrapolation, such a step was crucial.

The overall workflow of our approach is therefore to: 1) use the SKR API to extract semantic predications using SemRep; 2) build a predications graph from the SKR output, augmented with missing semantic predications we manually added for vascular reactivity; 3) use a DFS algorithm to traverse the predications graph to extract semantic associations between DFO and RS. The output of the algorithm is a set of reachability relations for each (DFO, RS) pair; 4) manually construct subgraphs by grouping associations deemed relevant to each of Swanson's original claims, using the collection of reachability relations from the DFS algorithm; 5) use domain expertise to add background knowledge to bridge knowledge gaps among concepts, in scenarios when subgraphs lack expressiveness.

Before discussing the experiments, we acknowledge the scalability limitation in manually clustering associations into subgraphs, and using domain experts to infer background knowledge. We anticipate in future work, the use of structural graph-based features such as centrality, geodesic and clustering coefficient [39] as well as semantics-based features such as the use of higher-order associations from ontology schemas [6, 37], concept class membership based on hierarchical and associative relationships [35], as well as semantic similarity measures, to improve scalability. Information-theoretic approaches such as information gain could also be useful. To support automatic inclusion of background knowledge, ideas such as knowledge abstraction we presented in [35] could also be pivotal. In the next Section we discuss the results of the two experiments conducted based on these steps.

4. Experimental Results

In the first experiment, we aim to show that Swanson's associations can be recovered and decomposed using semantic predications, semantic associations, expressive subgraphs, background knowledge and our graph-based techniques. Our results shows that we recovered the 3 primary associations, and retrieved 4 of the 8 of the decomposed supplementary associations and also 7 of the 8 of the decomposed secondary associations. An interesting observation in this process is that none of Swanson's original associations (primary or otherwise) occurred directly as individual associations with the reachability relations. It turns out that aggregating and clustering the associations was an imperative step in order to arrive at any of Swanson's conclusions.

Table 2: Semantic Association statistics

	B1:Full Text Articles					
Root	FOD		FO		EPA	
	382		2124		17848	
Terminal	RD	RP	RD	RP	RD	RP
	1	2	3	11	48	124

	B2:Titles/Abstracts					
Root	FOD		FO		EPA	
	22		67		224	
Terminal	RD	RP	RD	RP	RD	RP
	0	0	0	2	8	26

For the rest of this Section and the following Subsections, in each Figure solid black lines represent semantic predications that directly appeared in the literature. Solid double green lines represent assertions made by abductive reasoning using the surrounding semantic predications in the subgraph, and broken blue lines represent assertions gleaned from background knowledge. Also recall that Dietary Fish Oil (DFO) is used as reference for Fish Oils (FO), Fish Oil - dietary (FOD), Eicosapentaenoic Acid (EPA).

4.1. Experiment I: Titles, Abstracts and Full Text

For the first experiment (based on full text articles from B1), we found a total of 382 associations in the reachability relation, for which Fish Oil - dietary (FOD) was the root. Among these, only three associations terminated with RS (i.e., one with Raynaud Disease (RD=1) and two with Raynaud Phenomenon (RP=2) as shown in Table 2). We also found a total of 2124 associations for which Fish Oils (FO) was the root of the reachability relation. Among these only 14 terminated with RS (RD=3, RP=11). Then we found another 17848 associations for which Eicosapentaenoic Acid (EPA) was the root of the reachability relation, among which 172 terminated with RS (RD=48, RP=124). Hence, we found a total of 189 associations for which DFO was the root and RS was the terminal.

To link these automatically generated associations with Swanson's original associations, we manually selected relevant association from these 189 associations, and constructed subgraphs for comparison with each of the three primary associations (Table 3, ID:#1, #2 & #3), eight supplementary associations (Table 3, ID:1a, 2a-c & 3a-d) and eight secondary associations (Table5, ID:1.1-3, 2.1-2, 3.1-3) identified from Swanson's paper. We begin by comparing the subgraphs ($G_{P_i^{DFS}}$) and corresponding Swanson association (p_i^S) for platelet aggregation in the following Section.

4.1.1. Platelet Aggregation

The first primary association Swanson established (repeated in Table 4, ID:#1) was that (DFO INHIBITS platelet aggregation) and high levels of (platelet aggregation CAUSES RS). Among the 189 automatically generated associations, eight associations (Table 4, ID: 77,78,80,81,83,84,135&150) were deemed relevant to this claim. We constructed the subgraph in Figure 2 from these eight associations, and used it to conclude that since (DFO STIMULATES Epoprostenol) and Epoprostenol both DISRUPTS platelet aggregation and TREATS RS, then by abduction, perhaps (platelet aggregation CAUSES RS). If this is true, then (DFO TREATS RS) by stimulating Epoprostenol which disrupts platelet aggregation and thereby prevents RS.

From this example, it is clear that the integration of several associations, to form subgraphs is necessary for recovering Swanson's original associations. In this work, we have shown that the construction of such subgraphs can be semi-automated by automatically extracting the relevant semantic associations. We believe that the future of LBD will benefit from ability to fully automate the creation of such subgraphs.

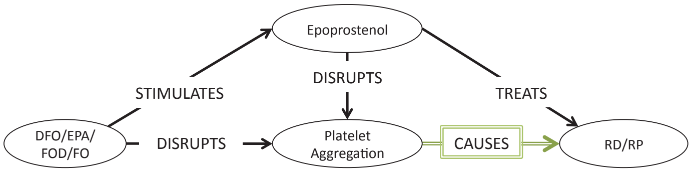

Figure 2: Primary Association Subgraph: Platelet Aggregation

Next, the only supplementary association involving DFO, RS, platelet aggregation (Table 3 & Table 4 ID:1a) that Swanson described, claims that (DFO PRODUCES Prostaglandin (PGI_3)) and (Prostaglandin (PGI_3) INHIBITS platelet aggregation) and (platelet aggregation CAUSES RS). Twelve associations (Table 4 ID:1a, Subgraph) were deemed relevant to this claim. From them, the subgraph in Figure 3 was constructed. This subgraph shows that (Eicosapentaenoic Acid CONVERTS_TO Prostaglandin (PGI_3)) and (Prostaglandin (PGI_3) ISA Epoprostenol). Since it was inferred from the previous subgraph in Figure 2 that (Epoprostenol TREATS RS) by disrupting platelet aggregation, by abduction we can surmise that Prostaglandin (PGI_3) also TREATS RS by disrupting platelet aggregation, because Prostaglandin (PGI_3) is also a Prostaglandin. Furthermore, since Epoprostenol is a member of the Prostaglandin family, we can also infer that (DFO TREATS RS) by producing Prostaglandin, which DISRUPTS platelet aggregation. This conclusion is supported by the first secondary association (Table 5, ID:1.1&73) in which Swanson claims that another Prostaglandin (PGE_1) (also called Alprostadil) also TREATS RS by inhibiting platelet aggregation.

Swanson also presented another secondary association (Table 5, ID:1.3) in which the drug Nifedipine, which ISA Calcium Channel Blocker, inhibits platelet activation and so TREATS RS. We found one association (Table 5, ID: 91) which states that (Nifedipine DISRUPTS Platelet function). We also found a second association (Table 5, ID:140) which states that (Nifedipine TREATS Raynaud Phenomenon). Since (Prostaglandin TREATS RS) by inhibiting platelet aggregation we can also surmise by abduction that Nifedipine also TREATS Raynaud Phenomenon as a result of disrupting platelet function. That (Nifedipine ISA Calcium Channel Blocker) would be obtained from background knowledge.

4.1.2. Blood Viscosity

The second primary association between DFO and RS (Table 3, ID:2), claims that (DFO INHIBITS blood viscosity) and (blood

Table 3: Primary & Supplementary Associations

Intermediate	ID	Association
Platelet Aggregation	#1	Dietary Fish Oil→INHIBITS→platelet aggregation→CAUSES→Raynaud Syndrome
	1a	Dietary Fish Oil→ PRODUCES→Prostaglandin (PGI_3)→INHIBITS→platelet aggregation→CAUSES→Raynaud Syndrome
blood viscosity	#2	Dietary Fish Oil→INHIBITS→blood viscosity→CAUSES →Raynaud Syndrome
	2a	Dietary Fish Oil→INHIBITS→triglyceride→ISA→Blood Lipid→AFFECTS→blood viscosity→CAUSES →Raynaud Syndrome
	2b	Dietary Fish Oil→AUGMENTS→Erythrocyte Deformability→INHIBITS→blood viscosity→CAUSES→Raynaud Syndrome
	2c	Dietary Fish Oil → INHIBITS → Serotonin → AUGMENTS → Erythrocyte Deformability → INHIBITS → blood viscosity → CAUSES → Raynaud Syndrome
vascular reactivity	#3	Dietary Fish Oil→INHIBITS→vascular reactivity→CAUSES→Raynaud Syndrome
	3a	Dietary Fish Oil→ PRODUCES→Prostaglandin (PGI_3)→CAUSES→vasodilation→INHIBITS→Raynaud Syndrome
	3b	Dietary Fish Oil→ PRODUCES→Prostaglandin (PGI_3)→INHIBITS→vasoconstriction→CAUSES→Raynaud Syndrome
	3c	Dietary Fish Oil → PRODUCES → Prostaglandin (PGE_1) → INHIBITS → platelet aggregation → CAUSES → vasoconstriction → CAUSES → Raynaud Syndrome
	3d	Dietary Fish Oil→INHIBITS→Serotonin→CAUSES→vasoconstriction→CAUSES→Raynaud Syndrome

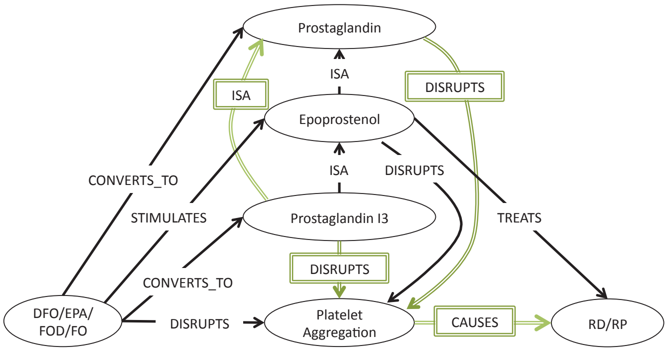

Figure 3: Supplementary Association Subgraph: Platelet Aggregation

viscosity CAUSES RS). We found four associations (Table 4, ID:71,135,150&175) deemed relevant to this claim, and from them constructed the subgraph in Figure 4.

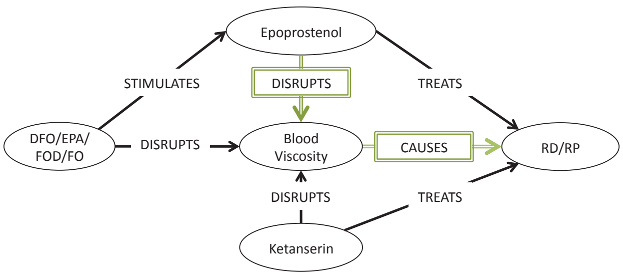

Figure 4: Primary Association Subgraph: Blood Viscosity

When we apply Hristovski's discovery pattern from [17] to this subgraph, we can surmise by adduction that if (Epoprostenol TREATS RS) and (Ketanserin TREATS RS), given that (Ketanserin DISRUPTS blood viscosity), then perhaps (Epoprostenol TREATS RS) by also disrupting blood viscosoty. If this is true, it follows that (DFO DISRUPTS blood viscosity) by producing Epoprostenol. We can therefore conjecture that blood viscosity may be another cause of RS.

Swanson's first supplementary association (Table 4, ID:2a) in which (DFO TREATS RS) through blood viscosity claims that DFO inhibits various blood lipids (specifically triglycerides) which directly or indirectly increase blood viscosity. We found six associations (Table 4, ID:71,111,112,135,150,175) deemed relevant to this observation and constructed the subgraph in Figure 5. From this subgraph, we observe that (DFO ISA Fatty Acid) and (DFO also DISRUPTS blood viscosity). However, from background knowledge [40], it is known that DFO is an essential fatty acid which is known to exhibit several health benefits. One major benefit of fatty acids is the inhibition of triglycerides (also a Lipid). This inhibition AFFECTS blood viscosity. It follows by abduction that (DFO TREATS RS) by inhibiting lipids and thereby lowering blood viscosity.

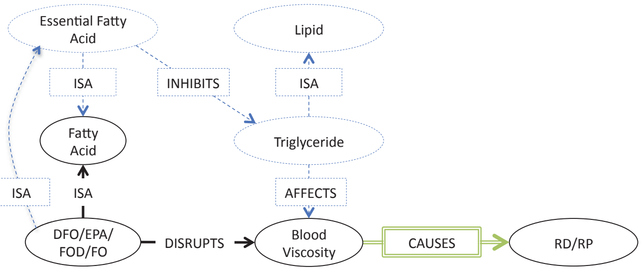

Figure 5: Supplementary Association Subgraph: Blood Viscosity

This example clearly establishes the need for background knowledge, whether explicit from an ontology or using domain experts to bridge the knowledge gap needed to make scientific discoveries from scientific literature. It highlights the third point put forth in the *Conclusion* Section of the abstract, which states that a critical aspect of semantics-based LBD will depend in future on *the semantic integration of background knowledge for interpretation*.

Swanson further discussed two additional supplementary as-

Table 4: Subgraphs for Primary and Supplementary Associations

Assoc.	#1	Dietary Fish Oil→INHIBITS→platelet aggregation→CAUSES→Raynaud Syndrome		
Subgraph	77	Eicosapentaenoic Acid DISRUPTS Platelet aggregation	Epoprostenol DISRUPTS Platelet aggregation	Epoprostenol TREATS Raynaud Disease Epoprostenol TREATS Raynaud Phenomenon
	78			
	80		Epoprostenol INHIBITS_MAN Platelet aggregation	
	81			
	83		Epoprostenol PREVENTS_MAN Platelet aggregation	
	84			
	135	Eicosapentaenoic Acid STIMULATES Epoprostenol		
	150			

Assoc.	1a	Dietary Fish Oil→ PRODUCES→Prostaglandin (PGI_3)→INHIBITS→platelet aggregation→CAUSES→Raynaud Syndrome		
Subgraph	37	Eicosapentaenoic Acid CONVERTS_TO prostaglandin I3	prostaglandin I3 ISA Epoprostenol	Epoprostenol TREATS Raynaud Disease Epoprostenol TREATS Raynaud Phenomenon
	38			
	40	Eicosapentaenoic Acid CONVERTS_TO Prostaglandins	Epoprostenol ISA Prostaglandins	
	41			
	77	Eicosapentaenoic Acid DISRUPTS Platelet aggregation	Epoprostenol DISRUPTS Platelet aggregation	
	78			
	80		Epoprostenol INHIBITS_MAN Platelet aggregation	
	81			
	83		Epoprostenol PREVENTS_MAN Platelet aggregation	
	84			
	135	Eicosapentaenoic Acid STIMULATES Epoprostenol		
	150			

Assoc.	#2	Dietary Fish Oil→INHIBITS→blood viscosity→CAUSES→Raynaud Syndrome		
Subgraph	71	Eicosapentaenoic Acid DISRUPTS blood viscosity	Ketanserin DISRUPTS blood viscosity	Ketanserin TREATS Raynaud Phenomenon
	175	Fish Oils AFFECTS blood viscosity		
	135	Eicosapentaenoic Acid STIMULATES Epoprostenol		Epoprostenol TREATS Raynaud Disease
	150			Epoprostenol TREATS Raynaud Phenomenon

Assoc.	2a	Dietary Fish Oil → INHIBITS → triglyceride → ISA → Blood Lipid → AFFECTS → blood viscosity → CAUSES → Raynaud Syndrome		
Subgraph	71	Eicosapentaenoic Acid DISRUPTS blood viscosity	Ketanserin DISRUPTS blood viscosity	Ketanserin TREATS Raynaud Phenomenon
	175	Fish Oils AFFECTS blood viscosity		
	135	Eicosapentaenoic Acid STIMULATES Epoprostenol		Epoprostenol TREATS Raynaud Disease Epoprostenol TREATS Raynaud Syndrome
	150			
	111	Eicosapentaenoic Acid ISA Fatty Acids	Epoprostenol STIMULATES Fatty Acids	
	112			
Assoc.	2b	Dietary Fish Oil → AUGMENTS → Erythrocyte Deformability → INHIBITS → blood viscosity → CAUSES → Raynaud Syndrome		
Assoc.	2c	Dietary Fish Oil → INHIBITS → Serotonin → AUGMENTS → Erythrocyte Deformability → INHIBITS → blood viscosity → CAUSES → Raynaud Syndrome		

Assoc.	#3	Dietary Fish Oil→INHIBITS→vascular reactivity→CAUSES→Raynaud Syndrome		
Subgraph	25	Eicosapentaenoic Acid AFFECTS_MAN Vascular constriction	Epoprostenol INHIBITS_MAN Vascular constriction	Epoprostenol TREATS Raynaud Disease
	26			Epoprostenol TREATS Raynaud Phenomenon

Assoc.	3a	Dietary Fish Oil → PRODUCES → Prostaglandin (PGI_3) → CAUSES → vasodilation → INHIBITS → Raynaud Phenomenon		
Assoc.	3b	Dietary Fish Oil → PRODUCES → Prostaglandin (PGI_3) → INHIBITS → vasoconstriction → CAUSES → Raynaud Phenomenon		
Subgraph	25	Eicosapentaenoic Acid AFFECTS_MAN Vascular constriction	Epoprostenol INHIBITS_MAN Vascular constriction	Epoprostenol TREATS Raynaud Disease Epoprostenol TREATS Raynaud Phenomenon
	26			
	66	Eicosapentaenoic Acid CONVERTS_TO prostaglandin I3	prostaglandin I3 ISA Epoprostenol	
	67			
	37	Eicosapentaenoic Acid CONVERTS_TO Prostaglandins	Epoprostenol ISA Prostaglandins	
	38			
Assoc.	3c	Dietary Fish Oil → PRODUCES → Prostaglandin (PGE_1) → INHIBITS → platelet aggregation → CAUSES → vasoconstriction → CAUSES → Raynaud Syndrome		
Assoc.	3d	Dietary Fish Oil → INHIBITS → Serotonin → CAUSES → vasoconstriction → CAUSES → Raynaud Syndrome		

sociations in which (DFO INHIBITS blood viscosity). In the first (Table 4, ID:2b), he claims that DFOs influence the ability of red blood cells to alter their shape under fluid pressure. That is, (DFO AUGMENTS Erythrocyte Deformability) and (Erythrocyte Deformability INHIBITS blood viscosity), and high (blood viscosity CAUSES RS). In the other supplementary association (Table 4, ID:2c) Swanson claims that (DFO INHIBITS Serotonin) and it is this inhibition that AUGMENTS Erythrocyte Deformability. Unfortunately, we did not recover any of these two associations.

Finally, Swanson also discussed another secondary association involving blood viscosity (Table 5, ID:2.2) in which (Ketanserin INHIBITS Serotonin) and hence AUGMENTS Erythrocyte Deformability. While this association was not recovered at this level of granularity, two associations (Table 5, ID:71,175) provided evidence that (Ketanserin TREATS RS) by inhibiting blood viscosity.

4.1.3. Vascular Reactivity

Swanson's third primary association (Table 3, Table 4, ID:#3) claims that (DFO TREATS RS) by inhibiting vascular reactivity (i.e., vasoconstriction, Vascular constriction (function)) which causes RS. We found two relevant associations (Table 4, ID:25,26) from which the subgraph shown in Figure 6 was constructed. This subgraph shows that (Eicosapentaenoic Acid AFFECTS_MAN Vascular constriction (function)) and (Epoprostenol INHIBITS_MAN Vascular constriction (function)) and (Epoprostenol TREATS RS). Since, we know that DFO STIMULATES Epoprostenol, and Epoprostenol also INHIBITS vasoconstriction and also TREATS RS, we can again conjecture by abduction that perhaps (vasoconstriction CAUSES RS).

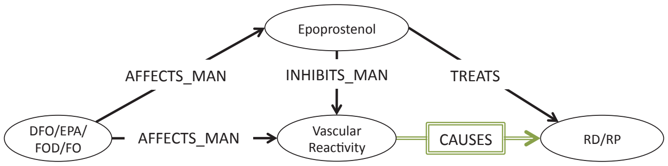

Figure 6: Primary Association Subgraph: Vascular Reactivity

The first supplementary association involving vasoconstriction (Table 4, ID:3a) states that (DFO PRODUCES PGI_3) and (PGI_3 CAUSES vasodilation), and vasodilation INHIBITS RS. The second supplementary association (Table 4, ID:3b) is similar, except that it states that (PGI_3 also INHIBITS vasoconstriction) and therefore INHIBITS RS. We found six associations (Table 4, ID:25,26,37,38,66,67) deemed relevant to this association. From them we surmized that perhaps (Epoprostenol TREATS RS) by inhibiting vasoconstriction. If (DFO PRODUCES Prostaglandin (PGI_3)) and (Prostaglandin (PGI_3) ISA Epoprostenol), given that (Epoprostenol INHIBITS vasoconstriction), Prostaglandin (PGI_3) possibly also INHIBITS vasoconstriction. Hence, it may be the case that (Epoprostenol TREATS RS) by inhibiting vasoconstriction.

We did not recover the third or fourth supplementary associations (Table 4, ID:3c,3d), hence we found no evidence that (Prostaglandin (PGE_1) CAUSES vasoconstriction) by inhibiting platelet aggregation.

Finally, Swanson also discussed several secondary associations involving vascular reactivity and RS. Again, we could not recover the secondary association in Table 5, ID:3.1, due the absence of predication which states that (Prostaglandin (PGE_1) CAUSES vasoconstriction).

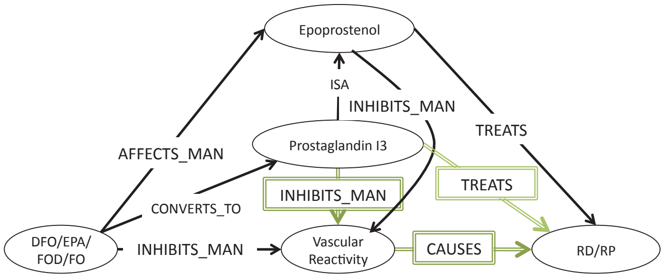

Figure 7: Supplementary Association Subgraph:Vascular Reactivity

For the secondary association in Table 5, ID:3.2, we found two associations Table 5, ID:25,26) from which we conjectured by abduction that since (Epoprostenol INHIBITS vasoconstriction) by inference (Epoprostenol CAUSES vasodilation) which (INHIBITS RS). For the last secondary association Table 5, ID:3.3), we found one association (Table 5, ID:35) from which we conjectured that since (Nifedipine INHIBITS vasoconstriction) it CAUSES vasodilation which INHIBITS RS. From background knowledge, it is known that (Nifedipine ISA Calcium Channel Blocker). In the next Section, we discuss the results for our second experiment which used only titles and abstracts.

4.2. Experiment II: Titles and Abstracts Only

In the second experiment, we aim determine to what extent titles and abstracts may be adequate for supporting LBD, if indeed that are adequate at all. After applying the DFS algorithm to the predications from the B2 dataset, we found 22 associations for which Fish Oil - dietary (FOD) was the root of the reachability relation. However no associations terminated with RS (RD=0,RP=0) as shown in Table 2. We also found a total of 67 associations for which Fish Oils (FO) was the root of the reachability relation, among which 2 terminated with RS (RD=0, RP=2). Finally, we also found another 224 associations, for which Eicosapentaenoic Acid (EPA) was the root of the reachability relation, among which 34 terminated with RS (RD=8, RP=26). Hence, we found a total of 36 associations for which DFO was the root and RS was the terminal. We again manually constructed the subgraphs and compared them with Swanson's original claims.

4.2.1. Platelet Aggregation

For Swanson's first primary association (Table 3, ID:1)), four associations (available online)[4] among these 36 associations were deemed relevant. We constructed the subgraph in Figure 8 from these associations, but no direct link between DFO and platelet aggregation was observed. Instead, two novel associations were found. Unfortunately, the absence of a direct link to platelet aggregation made it difficult to infer the role of DFO and platelet aggregation in treating RS. This indicates that the B2 dataset may not contain sufficient information to support recovery of Swanson's associations.

[4] Datasets - `wiki.knoesis.org/index.php/Obvio#RS-DFO_Hypothesis`

Table 5: Secondary Associations

Assoc.	1.1	Prostaglandin (PGE_1) → INHIBITS → platelet aggregation → CAUSES → Raynaud Syndrome		
Subgraph	73	Eicosapentaenoic Acid DISRUPTS Platelet aggregation	Alprostadil DISRUPTS Platelet aggregation	Alprostadil INTERACTS_WITH Raynaud Disease

Assoc.	1.2	Prostacyclin (PGI_2) → INHIBITS → platelet aggregation → CAUSES → Raynaud Syndrome		
Subgraph	77	Eicosapentaenoic Acid DISRUPTS Platelet aggregation	Epoprostenol DISRUPTS Platelet aggregation	Epoprostenol TREATS Raynaud Disease Epoprostenol TREATS Raynaud Phenomenon
	78			
	80		Epoprostenol INHIBITS_MAN Platelet aggregation	
	81			
	83		Epoprostenol PREVENTS_MAN Platelet aggregation	
	84			
	135	Eicosapentaenoic Acid STIMULATES Epoprostenol		
	150			

Assoc.	1.3	Nifedipine → ISA → CCB → INHIBITS → platelet activation → ASSOCIATED_WITH → Raynaud Syndrome		
Subgraph	91	Eicosapentaenoic Acid DISRUPTS Platelet function	Nifedipine DISRUPTS Platelet function	Nifedipine TREATS Raynaud Phenomenon
	140	Eicosapentaenoic Acid STIMULATES Epoprostenol	Epoprostenol TREATS Raynaud Phenomenon	

Assoc.	2.1	Prostacyclin (PGI_2) → AUGMENTS → Erythrocyte Deformability → INHIBITS → blood viscosity → CAUSES → Raynaud Syndrome (in vitro)		
Subgraph	71	Eicosapentaenoic Acid DISRUPTS blood viscosity	Ketanserin DISRUPTS blood viscosity	Ketanserin TREATS Raynaud Phenomenon
	175	Fish Oils AFFECTS blood viscosity		
	135	Eicosapentaenoic Acid STIMULATES Epoprostenol		Epoprostenol TREATS Raynaud Disease
	150			Epoprostenol TREATS Raynaud Phenomenon

Assoc.	2.2	Ketanserin → INHIBITS → serotonin → AUGMENTS → Erythrocyte Deformability → INHIBITS → blood viscosity → CAUSES → Raynaud Syndrome		
Subgraph	71	Eicosapentaenoic Acid DISRUPTS blood viscosity	Ketanserin DISRUPTS blood viscosity	Ketanserin TREATS Raynaud Phenomenon
	175	Fish Oils AFFECTS blood viscosity		

Assoc.	3.1	Prostaglandin (PGE_1) → CAUSES → vasodilation → INHIBITS → Raynaud Syndrome

Assoc.	3.2	Prostacyclin (PGI_2) → CAUSES → vasodilation → INHIBITS → Raynaud Syndrome		
Subgraph	25	Eicosapentaenoic Acid AFFECTS_MAN Vascular constriction	Epoprostenol INHIBITS_MAN Vascular constriction	Epoprostenol TREATS Raynaud Disease
	26			Epoprostenol TREATS Raynaud Phenomenon

Assoc.	3.3	Nifedipine → ISA → Calcium Channel Blocker → CAUSES → vasodilation → INHIBITS → Raynaud Syndrome		
Subgraph	35	Eicosapentaenoic Acid AFFECTS_MAN Vascular constriction	Nifedipine INHIBITS_MAN Vascular constriction	Nifedipine TREATS Raynaud Phenomenon

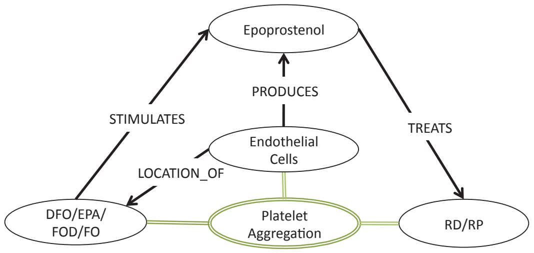

Figure 8: Primary Association Subgraph: Platelet Aggregation (Exp2)

For the first supplementary association (Table 3, Table 5, ID:1a) involving platelet aggregation, we found four of the eight associations from the first experiment. However, no direct association existed between DFO and platelet aggregation in the subgraph shown in Figure 9. While, we can conjecture that (DFO TREATS RS) through Prostaglandins from the two associations which state that (DFO PRODUCES Prostaglandin PGI_3) and (Epoprostenol ISA Prostaglandin), the role of platelet aggregation is still not apparent.

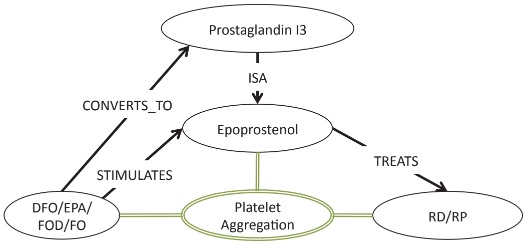

Figure 9: Supplementary Association Subgraph: Platelet Aggregation (Exp2)

Similarly for the secondary associations (Table 5, ID1.1,1,2,1,3), although there is sufficient information to surmize that (DFO TREATS RS), the role of platelet aggregation is again obscured.

4.2.2. *Blood Viscosity*

For Swanson's second primary association involving blood viscosity (Table 3, Table 5, ID: 2)), we found virtually the same four associations as in experiment 1 . The main difference was that for two of the four associations, we found that (Ketanserin AFFECTS blood viscosity), instead of (Ketanserin DISRUPTS blood viscosity). From the subgraph in Figure 10, we can conclude that (DFO TREATS RS) by lowering blood viscosity, on the assumption that (Ketanserin AFFECTS blood viscosity) is synonymous with (Ketanserin DISRUPTS blood viscosity) and by applying Hristovski's discovery pattern to infer that Epoprostenol also DISRUPTS blood viscosity.

By contrast however, for Swanson's first supplementary association (Table 4, ID:2a), we recovered four associations out of six associations from experiment 1. The critical association between DFO and fatty acids which states that (Eicosapentaenoic Acid ISA Fatty Acids) and (Epoprostenol STIMULATES Fatty Acids) are notably absent. It was difficult to make the association that (DFO TREATS RS) by inhibiting blood viscosity.

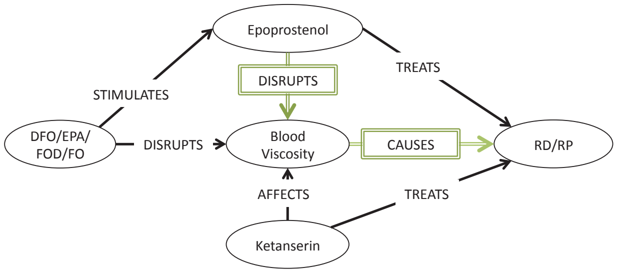

Figure 10: Primary Association Subgraph: Blood Viscosity (Exp2)

4.2.3. *Vascular Reactivity*

There were two distinct mentions of vascular reactivity in the B2 dataset, from which one predication was manually identified, that is, (vasoconstriction ASSOCIATED_WITH Eicosapentaenoic Acid). Unfortunately this was insufficient for linking DFO and RS through vascular reactivity. Hence, no associations were recovered using the B2 dataset.

This result seems to suggests that titles and abstracts alone, might NOT be sufficient to rediscover Swanson's hypothesis, hence raising concerns about their adequacy for Literature-Based Discovery (LBD) in general.

5. Discussion

Our graph-based framework leverages semantic representations to recover and decompose Swanson's *Raynaud Syndrome–Fish Oil* Hypothesis. To the best of our knowledge, this is the first attempt at decomposing Swanson's primary associations into supplementary and secondary associations (shown in Table 6). We therefore exceed the state-of-the-art in semantics-based approaches for recovering Swanson's *RS-DFO* Hypothesis [17, 30] in terms of context, coverage, and expressiveness of associations, while providing insight into some prerequisites for semantics-based LBD.

However, while an improvement over the state-of-the-art, our graph and semantics-based approach is far from foolproof. This anecdotal example using Swanson's *RS-DFO* Hypothesis, circumvents many limitations that exist when using large corpora and background knowledge sources for open discovery.

5.1. *Limitations*

The first limitation is the availability of the relevant datasets that contain adequate information to support LBD. While MEDLINE provides titles and abstracts for more than 20 million scientific articles, it is unclear to what extent this can be exploited for LBD. Experiment 2 shows compelling evidence that many of the associations recovered in Experiment 1, which were critical for reasoning were unavailable in the titles and abstracts in the second experiment. This becomes a major problem when addressing open discovery, given the full text articles may be unavailable. The concern is that *"hidden information may remain hidden"* in full text throughout the corpus.

Table 6: Comparison of Semantics-based techniques for recovering and decomposing Swanson's Associations

The second limitation is the ability to extract semantic information from the selected literature [31]. As evidenced by the vascular reactivity scenario, the inability to extract some semantic predications presents a major bottleneck. While we avoided this problem owing to manual annotations on this small dataset, semantic information extraction becomes a serious problem on the large scale.

The third issue is devising techniques to avoid the combinatorial explosion that arises when traversing large data graphs. Many query execution platforms will time-out, due to memory limitations, and return no results on large graphs. While, graph traversal can be propagated based on as structural graph properties [39], the need to exploit semantics is crucial. Semantics-based techniques for addressing combinatorial explosion have been proposed [41, 35], but their effectiveness for open discovery scenarios has not yet been realized.

Furthermore, another issue is deriving scalable techniques for extracting relevant semantic associations between concepts, and then aggregating such associations to form subgraphs. Subgraph creation is crucial for sense making, question answering and ultimately LBD. While we anticipate that the extraction of semantic associations can be based on the notion of ρ-*isomorphism* [6] and information theory [42], semantic association extraction from large data graph is still a major problem. Popular pattern-based approaches [17, 30, 19] are not scalable, while structural graph-based approaches [16, 28, 20] tend to lack both coverage and expressiveness.

Beyond the extraction of semantic associations from large data graphs and the automatic creation of subgraphs from those associations, perhaps the most significant challenge is incorporating background knowledge from structured sources to enhance subgraphs to bridge knowledge gaps when subgraphs themselves prove insufficient. Aside from work by Cameron et. al. [35], the semantic integration of scientific literature and background is considered next generation research and thus remains largely unexplored. Although Russ et. al. [43] developed KEfED for the semantic integration of background knowledge and other knowledge sources, the focus is primarily on experimental data rather than scientific literature.

Finally, another major issue is resolving inconsistencies arising during the semantic predication extraction. For instance, both (Eicosapentaenoic Acid INHIBITS Epoprostenol) and (Eicosapentaenoic Acid STIMULATES Epoprostenol) were observed from the SemRep output. Such discrepancies must be resolved at some level, if a robust graph-based semantic system for LBD can be created.

5.2. Contribution

In spite of these limitations, we showed that semantic predications, semantic associations, subgraphs, background knowledge and our graph-based techniques are important elements for LBD. Our results advance the standard established by the state-of-the-art when recovering and decomposing Swanson's *RS-DFO* Hypothesis using these constructs. This suggests that a critical feature of next generation LBD is the semantic integration of scientific literature with background knowledge.

5.3. Implications

A fully mature system implemented based on our methodology, could provide the following features. Through a web interface, a user may select a concept from the predications graph, then browse intuitively along some association. After reaching a suitable target, select an option to generate the subgraph of associations related to the current association. Concurrently, the system will present the related predications from the background knowledge source that logically connect disjoint concepts within the generated subgraph. Such functionalities we believe will be significant in aiding LBD.

6. Conclusion

We presented a graph-based methodology for LBD that uses semantic predications, semantic associations, expressive subgraphs and background knowledge to recover and decompose Swanson's *Raynaud Syndrome-Fish Oil* hypothesis. We are the first to put into perspective the shift in critical constructs for LBD, from term co-occurrence and the ABC model, to semantic predications, semantic associations (using the *AnC model*), subgraphs and background knowledge. We are also the first to conduct an in-depth decomposition of Swanson's hypothesis based on this shift resulting in the retrieval of 14 out of the 19 associations, including 3 of the 3 primary associations, 4 of the 8 supplementary associations and 7 of the 8 secondary associations. In this process, we established that the use of graph-based techniques together with semantic representations to recover and decompose Swanson's hypothesis as well as to support LBD in general could benefit from: 1) expressive semantic representations beyond Swanson's ABC model; 2) an ability to accurately extract semantic information from text; and 3) an ability to semantically integrate background knowledge to facilitate interpretation.

7. Acknowledgement

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM) and the Kno.e.sis Center at Wright State University. We would like to thank Dongwook Shin and the ORISE program for their support, and Cartic Ramakrishnan, Pablo N. Mendes and Ramakanth Kavuluru for their ideas.

References

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

[2] Zahavi J, Hamilton W, O'Reilly M, Leyton J, Cotton L, Kakkar V. Plasma exchange and platelet function in raynauds phenomenon. Thromb Res 1980;19(1–2):85–93.

[3] Moncada S. Biology and therapeutic potential of prostacyclin. Stroke; a journal of cerebral circulation 1983;14(2):157–68.

[4] Moncada S, Vane JR. Prostacyclin and its clinical applications. Annals of clinical research 1984;16(5–6):241–52.

[5] Swanson DR. Undiscovered public knowledge. Library Quarterly 1986;56(1):103–18.

[6] Anyanwu K, Sheth AP. The ρ operator: Discovering and ranking associations on the semantic web. SIGMOD Record 2002;31(4):42–7.

[7] Swanson DR. Migraine and magnesium: eleven neglected connections. Perspect Biol Med 1988;31(4):526–57.

[8] Swanson DR. Somatomedin C and arginine: implicit connections between mutually isolated literatures. Perspectives in biology and medicine 1990;33(2):157–86.

[9] Smalheiser NR, Swanson DR. Indomethacin and Alzheimer's disease. Neurology 1996;46(2).

[10] Smalheiser NR, Swanson DR. Linking estrogen to Alzheimer's disease: an informatics approach. Neurology 1996;47(3):809–10.

[11] Smalheiser NR, Swanson DR. Calcium-independent phospholipase a2 and schizophrenia. Arch Gen Psychiatry 1998;55(8):752–3.

[12] Lindsay RK, Gordon MD. Literature-based discovery by lexical statistics. JASIS 1999;50(7):574–87.

[13] 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. JASIS 1996;47(2):116–28.

[14] Weeber M, Klein H, de Jong-van den Berg LTW, Vos R. Using concepts in literature-based discovery: Simulating swanson's raynaud-fish oil and migraine-magnesium discoveries. JASIST 2001;52(7):548–57.

[15] Srinivasan P. Text mining: Generating hypotheses from medline. JASIST 2004;55(5):396–413.

[16] Pratt W, Yildiz MY. LitLinker: capturing connections across the biomedical literature. In: Proceedings of the 2nd international conference on Knowledge capture. K-CAP '03; NY, USA: ACM; 2003, p. 105–12.

[17] Hristovski D, Friedman C, Rindflesch TC, Peterlin B. Exploiting semantic relations for literature-based discovery. Annual Symposium proceedings AMIA Symposium 2006;:349–53.

[18] Hristovski D, Peterlin B, Mitchell JA, Humphrey SM. Using literature-based discovery to identify disease candidate genes. I J Medical Informatics 2005;74(2-4):289–98.

[19] Ahlers CB, Hristovski D, Kilicoglu H, Rindflesch TC. Using the literature-based discovery paradigm to investigate drug mechanisms. AMIA Annu Symp Proc 2007;:6–10.

[20] Wilkowski B, Fiszman M, Miller C, Hristovski D, Arabandi S, Rosemblat G, et al. Discovery browsing with semantic predications and graph theory. AMIA Annu Symp Proc 2011;.

[21] Srinivasan P, Libbus B. Mining medline for implicit links between dietary substances and diseases. Bioinformatics 2004;20:290–6.

[22] Ganiz MC, Pottenger WM, Janneck CD. Recent advances in literature based discovery. Journal of the American Society for Information Science and Technology, JASIST 2006;.

[23] Swanson DR, Smalheiser NR. An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artificial Intelligence 1997;91(2):183 – 203.

[24] Smalheiser NR, Swanson DR. Using arrowsmith: A computer-assisted approach to formulating and assessing scientific hypotheses. Computer Methods and Programs in Biomedicine 1998;57(3):149 –53.

[25] Srinivasan P. Text mining: Generating hypotheses from medline. Journal of the American Society for Information Science and Technology 2004;55:396–413.

[26] Smalheiser NR. Literature-based discovery: Beyond the abcs. Journal of the American Society for Information Science and Technology 2012;63(2):218–24.

[27] Gordon MD, Dumais ST. Using latent semantic indexing for literature based discovery. JASIS 1998;49(8):674–85.

[28] Yetisgen-Yildiz M, Pratt W. Using statistical and knowledge-based approaches for literature-based discovery. Journal of Biomedical Informatics 2006;39(6):600–11.

[29] Cohen T, Whitfield GK, Schvaneveldt RW, Mukund K, Rindflesch T. EpiphaNet: An Interactive Tool to Support Biomedical Discoveries. Journal of biomedical discovery and collaboration 2010;5:21–49.

[30] Hristovski D, Friedman C, Rindflesch TC, Peterlin B. Literature-Based Knowledge Discovery using Natural Language Processing. In: Bruza P, Weeber M, editors. Literature-based Discovery; vol. 15 of *Information Science and Knowledge Management*; chap. 9. Berlin, Heidelberg: Springer Berlin Heidelberg; 2008, p. 133–52.

[31] Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: Interpreting hypernymic propositions in biomedical text. Journal of Biomedical Informatics 2003;36(6):462–77.

[32] Lussier Y, Borlawsky T, Rappaport D, Liu Y, Friedman C. Phenogo: assigning phenotypic context to gene ontology annotations with natural language processing. In: Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing. 2006,.

[33] Cameron D, Mendes PN, Sheth AP, Chan V. Semantics-empowered text exploration for knowledge discovery. In: 48th ACM Southeast Conference. 2010,.

[34] Kavuluru R, Thomas C, Sheth A, Chan V, Wang W, Smith A, et al. An up-to-date knowledge-based literature search and exploration framework for focused bioscience domains. In: 2nd ACM SIGHIT International Health Informatics Symposium. 2012,.

[35] Cameron D, Kavuluru R, Bodenreider O, Mendes PN, Sheth AP, Thirunarayan K. Semantic predications for complex information needs in biomedical literature. In: BIBM. 2011, p. 512–9.

[36] Tesseract . http://code.google.com/p/tesseract-ocr/. 2012.

[37] Anyanwu K, Sheth AP. ρ-queries: enabling querying for semantic associations on the semantic web. In: WWW. 2003, p. 690–9.

[38] Wikipedia . http://en.wikipedia.org/wiki/reachability. 2012.

[39] Elmacioglu E. On six degrees of separation in dblp-db and more. ACM SIGMOD Record 2005;34:33–40.

[40] Moghadasian MH. Advances in dietary enrichment with n-3 fatty acids. Critical Reviews in Food Science and Nutrition 2008;48(5):402–10.

[41] Anyanwu K, Maduko A, Sheth AP. Sparq2l: towards support for subgraph extraction queries in rdf databases. In: WWW. 2007, p. 797–806.

[42] Anyanwu K, Maduko A, Sheth AP. Semrank: ranking complex relationship search results on the semantic web. In: WWW. 2005, p. 117–27.

[43] Russ TA, Ramakrishnan C, Hovy EH, Bota M, Burns GAPC. Knowledge engineering tools for reasoning with scientific observations and interpretations: a neural connectivity use case. BMC Bioinformatics 2011;12:351.