ELSEVIER

Artificial Intelligence 91 (1997) 183–203

Artificial Intelligence

An interactive system for finding complementary literatures: a stimulus to scientific discovery

Don R. Swanson [a,*], Neil R. Smalheiser [b]

[a] *Division of the Humanities, The University of Chicago, 1010 E. 59th St., Chicago, IL 60637, USA*
[b] *Department of Pediatrics, MC 5058, The University of Chicago, 5841 S. Maryland Ave., Chicago, IL 60637, USA*

Received December 1995; revised June 1996

Abstract

An unintended consequence of specialization in science is poor communication across specialties. Information developed in one area of research may be of value in another without anyone becoming aware of the fact. We describe and evaluate interactive software and database search strategies that facilitate the discovery of previously unknown cross specialty information of scientific interest. The user begins by searching MEDLINE for article titles that identify a problem or topic of interest. From downloaded titles the software constructs input for additional database searches and produces a series of heuristic aids that help the user select a second set of articles complementary to the first set and from a different area of research. The two sets are complementary if together they can reveal new useful information that cannot be inferred from either set alone. The software output further helps the user identify the new information and derive from it a novel testable hypothesis. We report several successful tests and applications of the system. © 1997 Elsevier Science B.V.

1. Introduction and background

An important problem in the growth of knowledge is brought to light by the following type of literature structure: one set of articles (AB) reports an interesting association between variables A and B, a different set of articles (BC) reports a relationship between B and C, but nothing at all has been published concerning a possible link between A and C, even though such a link if validated would be of

* Corresponding author. E-mail: swanson@kiwi.uchicago.edu.

PII S0004-3702(97)00008-8

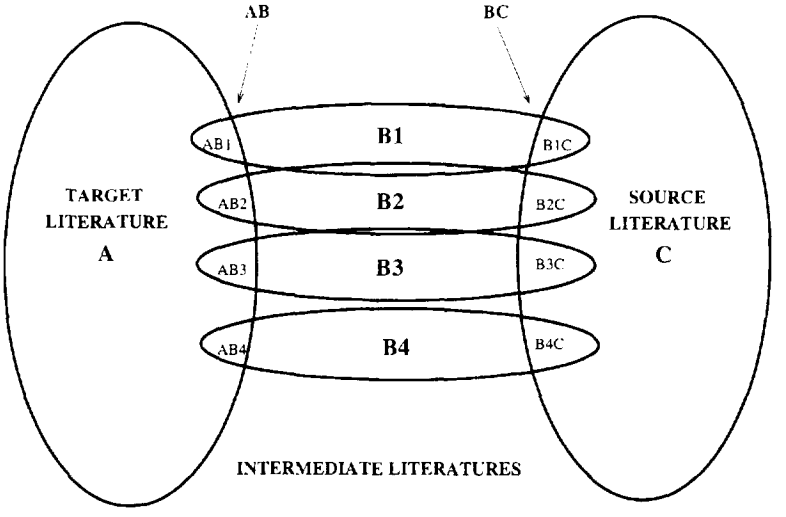

Fig. 1. A Venn diagram that represents sets of articles, or "literatures", A and C, that have no articles in common, but which are linked through intermediate literatures, Bi ($i = 1, 2 \ldots$). Such a structure may contain unnoticed useful information that can be inferred by combining pairs of intersections ABi and BiC.

scientific interest [6,30]. [1] For example, "A" might represent intake of a substance that induces a specific physiological change B, which in turn influences disease or body organ C. In broader terms, two literatures such as AB and BC are complementary if useful information can be inferred by considering them together that cannot be inferred from either one alone. If, furthermore, the two literatures are mutually isolated in that the authors and readers of one literature are not acquainted with the other, and vice versa, as may often be the case for two different specialties, then no one at all may be in a position to notice the implication that A and C might be related. To detect literature pairs that are mutually isolated, we examine the citation pattern [7]. If two literatures are "noninteractive"—that is, if they have never (or seldom) been cited together, and if neither cites the other—then it is possible that scientists have not previously considered both literatures together [25,35]. The two conditions, complementarity and noninteraction, describe a model structure that shows how useful information can remain undiscovered even though its components consist of public knowledge. Fig. 1 illustrates how multiple intermediate literatures might link a given A and C.

The concept of "undiscovered public knowledge" was developed and exemplified in 1986 [28,29] based on an actual structure in the biomedical literature similar to the AB–BC model shown in Fig. 1. Articles on Raynaud's disease (C) and articles on eicosapentaenoic acid (A) when considered together were suggestive that Raynaud patients might benefit from dietary fish oils rich in eicosapentaenoic acid [29]. One

[1] A, B, and C refer either to physiological conditions or agents or to the search terms based on them, or to the literatures that result from the search, the context indicating which meaning is intended. "B" may also stand, generically, for Bi with $i = 1, 2, 3, \ldots$, as in Fig. 1.

B-linkage, for example, was: dietary eicosapentaenoic acid can decrease *blood viscosity* (*B*); abnormally high *blood viscosity* has been reported in patients with Raynaud's disease. Neither of the two literatures (*A*, *C*) cited or mentioned the other, nor has extensive searching turned up any prior suggestions that dietary fish oils might influence Raynaud's disease. An independent clinical trial by medical researchers [2] subsequently corroborated the predicted beneficial effects, as discussed in [5,37]. This first structure was found more or less by accident, but inspired an ongoing effort to systematize the search process [32,36], corroborate and replicate the structure [8], model it [3,36], extend it [3,4,8], trace historical antecedents [4], and examine citations to the project itself [27,37].

A second example of complementary noninteractive literatures began with the literature on migraine (*C*), and subsequently identified a literature showing that magnesium deficiency (*A*) led to certain physiological effects (*B*) that, in a different context, were associated with migraine [31]. Eleven intermediate effects, *B*, provided the linkage. (Two such linkages are, for example: magnesium can inhibit *spreading depression* in the cortex, and *spreading depression* may be implicated in migraine attacks; magnesium deficient rats have been used as a model of *epilepsy*, and *epilepsy* has been associated with migraine). Remarkably, even though the literatures on magnesium and migraine that were identified as having extensive indirect linkages consisted of more than 60 articles each, neither literature cited or mentioned the other. Moreover at that time (1988) there were almost no records in the entire MEDLINE database that contained both of the words "migraine" and "magnesium". Since 1988 [31], more than 12 different groups of medical researchers have reported a systemic or local magnesium deficiency in migraine or a favorable response of migraine patients to dietary supplementation with magnesium [21,37]. There is also one recent report of negative results [26].

To date we have found and described seven examples of complementary pairs of literatures, using progressively improved methods for finding them [21–24,29,31,33]. In each case we discerned a novel testable hypothesis that was implicit in a pair of literatures considered together, but not previously made explicit in published form. Each structure was found through extensive searching, exploring, and reading biomedical literature [34]. The hypotheses evoked by these procedures can be evaluated by: (1) their inherent plausibility, (2) their acceptance for publication in a refereed biomedical journal; (3) whether they stimulate biomedical researchers to conduct clinical trials or laboratory experiments, and (4) whether the hypotheses are corroborated as a result of such tests. For the two earliest cases [29,31] all four criteria appear to have been met [2,21,37].

Our goal has been to create interactive software and database search strategies that can facilitate the discovery of complementary structures in the published literature of science. The universe of literature or search space under consideration is limited only by the coverage of the major scientific databases, though we have focused primarily on the biomedical field and the MEDLINE database (8 million records). In 1991, a systematic approach to finding complementary structures was outlined and became a point of departure for software development [36]. The system that has now taken shape is based on a 3-way interaction between computer software, bibliographic databases, and a

human operator. The interaction generates information structures that are used heuristically to guide the search for promising complementary literatures. In this paper we describe and evaluate the experimental computer software, which we call ARROWSMITH, and we explain how it functions within the system.

2. Approach and overview

For the migraine study [31] we analyzed the form of the eleven relationships between A and B and between B and C that we had found [36, p. 282]. (For example, some variant of "can or might influence" occurred repeatedly.) It was clear from the outset that, even if we were able to fit most relationships into a limited number of seemingly simple patterns, few or none of them were transitive; they were, however, often suggestive of a plausible new inference. But to recognize relationships within the natural language text of titles and abstracts and to draw inferences from them requires both common sense and extensive background knowledge. We have not attempted to formalize or automate such tasks. Our approach is based instead on automated procedures that present, for human observation, suggestive juxtapositions of natural language extracts (mainly titles, at present) from database records. We attempt to stimulate, rather than codify, the discovery process.

The user of the system begins by choosing a question or problem area of scientific interest that can be associated with a literature, C. The ultimate target of the ensuing search is a second literature, A, complementary to literature C. Two main procedures (I and II) implement the foregoing approach in a working experimental system. *Procedure* I operates on a single specified "source" literature, C, identified by a database search of title words or phrases. Fig. 2 is a schematic diagram of title word pathways from literature C, proceeding via multiple intermediate B-terms that co-occur with title word "C", to various possible target title words denoted $A1$, $A2$, $A3\ldots$, etc., each of which co-occurs with one or more B-terms. There can be, potentially, thousands of intermediate B-terms that link to C, and, for each B, thousands of possible candidates for A that link to B. The problem of searching for one (or a few) initially unknown Ai within the entire MEDLINE database presents a formidable explosion of possibilities. The interactive system we describe evades this explosion by first restricting, by various means, the number of different B-linkages to be traversed. From the resulting list of "A-candidates", the user chooses one Ai judged to be biologically plausible. (Alternatively, the user may choose A on some other basis and simply bypass *Procedure* I.)

Procedure II starts anew from *two* pre-selected literatures, A and C (irrespective of how A was chosen). Focusing on a single target literature, A, makes it feasible now to develop a much less restricted list of title word pathways that connect A to C. This procedure aims to identify all title word pathways that might provide clues to the presence of complementary arguments within those literatures. The output of *Procedure* II, a structured title display (plus journal citation), serves as a heuristic aid to identifying word linked titles and serves as an organized guide to the literature.

ARROWSMITH (and its user) may be seen as a problem generating system [16,17]. The user chooses a relatively broad initial problem area (e.g. a disease for which neither

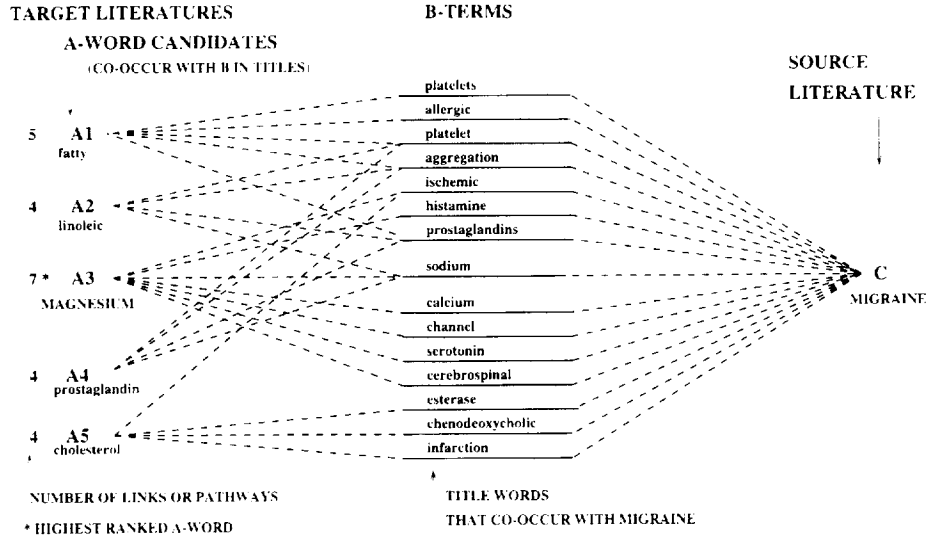

Fig. 2. A schematic diagram of title word pathways that, proceeding from right to left, lead from a source literature C through intermediate title words (B-terms) or literatures, to one or more title words that can represent promising target literatures, Ai ($i = 1, 2, \ldots$). A ranking algorithm for each Ai, based on the total number of B-pathways that lead to it, is illustrated. The rank numbers (number of paths) are shown in the leftmost column.

cause nor cure is known) and is guided by *Procedure* I toward promising complementary literatures and by *Procedure* II toward a series of more specific problems concerning biological pathways that might constitute mechanisms of action, and ultimately toward a problem presented to the experimentalist in the form of a plausible testable hypothesis.

2.1. *Procedure I: forming a ranked list of A-word candidates*

In this section we describe *Procedure* I applied to a specific example in which C is taken as the pre-1988 literature on migraine.

Step 1. Conduct a MEDLINE search for all titles that contain the word "migraine"; download these titles into a local computer file (FILE C). Derive automatically from FILE C a list of the unique words that it contains (that is, words that co-occur in titles with "migraine"). This list contains all possible candidates for B-words (including potentially interesting words such as "inflammatory", "platelet", "prostaglandin", "serotonin", "vasospasm") but it contains also many unsuitable words. In order to evade an explosion of search paths and to limit the output to manageable size, four types of restrictions are introduced, the first being: (i) words that are predictably unsuitable as B-candidates are excluded (e.g. not only nontopic words such as "able", "about", "above", but also words that are vague or too general such as "clinical", "comparative", "drugs", "evidence", "experimental", "role", "treatment", "studies") [36]. A pre-compiled exclusion list, or "stoplist" is provided to the computer. The list is

human constructed on the basis of judgment applied *a priori* concerning the suitability of each word, and at present consists of about 5000 words.

Step 2. Each of the remaining *B*-word candidates is then searched in MEDLINE to determine the total number of titles in which it occurs. Restriction (ii): these words are further screened automatically to retain only those that occur with greater relative frequency in migraine titles than in titles from MEDLINE as a whole. The latter frequency is determined from the information displayed in the online search which shows each search statement and the corresponding number of items found. More specifically, we retain only words for which the probability is small that a random allocation of words to titles could lead to a number of co-occurrences with "migraine" equal to or greater than the observed number [1]. (The cutoff probability in our migraine example is 0.09, as calculated from a Poisson distribution, ignoring multiple occurrences of a given word within the same title.)

Step 3. We introduce restriction (iii): the human operator examines the filtered list of *B*-terms, and removes entries that are judged not suitable—e.g. words that should be put on the stoplist, or possibly words too broad to be useful in the particular problem considered. Each word that remains becomes the basis for a new MEDLINE title search.

Restriction (iv) is introduced into the search strategy: each set of records formed by searching titles for each *B*-word is to be narrowed down to certain *categories* of records (chosen in advance) that are likely to contain target words of particular interest. This restriction is methodologically important for it provides the user flexibility in conducting an exploratory process that may in many respects model the process of scientific inquiry [11]. To date, most of our restriction categories have focused on exogenous agents because of their potential for experimental manipulation in clinical tests. Thus, categories such as deficiencies, dietary factors, toxins, or various subcategories of drugs or environmental factors might be selected, strategies that can be designed and tested in advance and built into the system [36]. In the present migraine example, the strategy was designed to restrict the output to dietary factors or deficiencies induced by dietary deprivation. Detailed search statements are given in [36]. Commands to form the intersection of the restriction set with each *B*-word set and a command to form the union of those intersections (U), are constructed automatically.

Step 4. The resulting search is uploaded into MEDLINE and executed. All titles that correspond to set U are then downloaded to FILE U in the local computer. FILE U is converted to a list of word occurrences (potential *A*-words), again using the stoplist as a filter, repeating restriction (i). Each *A*-occurrence is attached to the *B*-word that generated it, thus forming an *A*–*B* list that is used at a later point. It is also used immediately to create a list of unique *A*-words for the database search of Step 5.

Step 5. A database title search is conducted for each unique *A*-word in order to determine, for each *A*, the number of *A*-word titles within the restriction category. Repeating restriction (ii): the *A*-words are filtered according to probabilistic criteria, but now based on their co-occurrences with each *B*-word (rather than with "migraine"). The *A*-words thus selected constitute the list of *A*-candidates. Each candidate is then assigned a rank according to the number of different *B*-words in the *AB*–*BC* co-occurrence linkages in which it participates, as illustrated in Fig. 2. This ranking algorithm is based on a presumption that the greater the number of *B*-term linkages, the greater the

MIGRAINE	RAYNAUD'S DISEASE	ALZHEIMER'S DISEASE
7 magnesium	6 fatty	17 endotoxin
6 hormone	5 vitamin	13 toxin
6 b	4 sodium	13 methyl
5 salt	4 prostaglandins	13 d
5 pyridoxine	4 oil	12 lead
5 pressure	4 iga	12 ethanol
5 lipids	4 c	11 ozone
5 lipid	4 a2	11 necrosis
5 hypertension	4 5	9 neurotoxicity
5 hepatic	3 renin	9 calcium
5 fatty	3 polyunsaturated	9 c
5 e	3 platelets	8 seizures
5 d	3 magnesium	8 pertussis
5 cholesterol	3 lipid	8 mptp
4 vitamin	3 linoleic	8 hepatic
4 renin	3 glutathione	8 dna
4 prostaglandin	3 choline	7 tnf
4 methionine	3 angiotensin	7 il
4 membrane	3 anemia	7 b
4 linoleic		6 cadmium
4 folate		
4 c		
3 thiamine		
3 potassium		

Fig. 3. Top twenty (approx.) entries on the *A*-candidate lists for migraine, Raynaud's disease, and Alzheimer's disease, produced by *Procedure* I (Fig. 3). The ranking numbers shown in the first column represent the number of title word pathways connecting *A* to *C*, and were generated automatically by the process illustrated in Fig. 2. Not shown, but included in the actual output, are the *B*-terms (illustrated in Fig. 2) that contribute to the rank score.

chance that some of them will be biologically important. The candidate list is displayed for the human operator to use as a heuristic device (Fig. 3).

Step 6. From the ranked list, expert judgment is brought to bear in choosing words that appear to be the best candidates for assisting the discovery of complementary arguments, and so for defining a literature (*A*) complementary to that of *C*. Any one choice of *A*-word from the candidate list can provide input to *Procedure* II (Section 2.2).

The leftmost list in Fig. 3 shows the first 24 ranked *A*-terms for our migraine example. "Magnesium" is ranked first. Not only is it of high rank, it is biologically highly plausible. Magnesium is an essential element in the human diet and is well known as a modulator of neurotransmission. By these criteria "magnesium" emerged as the most promising *A*-word candidate.

Other words on the *A*-candidate list also might merit further exploration. Some of them are, however, too broad to be useful if taken alone ("hormone", "pressure", "lipid(s)", "hepatic", "membrane"). For reasons of practicality, the *A*-list is at present limited to single words, even though phrases may often be more appropriate and informative. However, the titles in which any particular *A*-word occurs (in FILE *U*) can be examined to determine whether a consistently more specific context can be identified. For example, we found that "hormone" and its *B*-linkages referred in most cases to

"growth hormone"; the *B*-linkages, however, did not appear to be important enough to pursue further. One could determine also whether a single letter such as "*e*" refers consistently to vitamin E, apolipoprotein E, E-coli, or hepatitis E, and so on.

2.2. *Procedure II: forming a list of B-terms and a title display*

This procedure develops a greatly expanded list of *B*-terms for a given *A* and *C*. The probabilistic and category restrictions ((ii) and (iv)) are dropped, and phrases of up to 6 contiguous words are included. "*A*" represents any single agent which may have been taken from the *Procedure* I list of *A*-candidates, or "*A*" may simply be based on an informed scientific conjecture concerning a possible *A*–*C* relationship.

Procedure II should always be preceded by a database search for all records that contain both *A* and *C* (not restricting the search to just titles), in order to identify any *A*–*C* or *A*–*B*–*C* relationships that are already explicitly published. Such known *B*-linkages should be investigated in advance of executing *Procedure* II to avoid unknowingly rediscovering them as indirect linkages; *Procedure* II can then focus on relationships that are either novel or at least cannot be found by conventional searching. Understanding strategies used in conventional searching is important at this point; finding the "direct" literature is not always straightforward [12,13]. The citation interaction pattern also can play a key role in determining whether an *A*–*C* relationship exists and is already known, a process described in more detail elsewhere [35].

The output of *Procedure* II is to be a printed display of titles from the *A* and *C* literatures organized according to the words, *B*, that they have in common. Fig. 4 is a schematic flowchart of the process that leads to this output, a process that begins with downloading the two sets of titles (and journal citations) for literatures *A* and *C*. For the migraine application (based on pre-1988 literature), MEDLINE was used to create a local file (C) of all titles (about 2800) that contain the word "migraine", and a file (A) of all titles (about 8000) that contain the word "magnesium". No title contained both words. The computer then produced a list of all words and phrases (about 200, after excluding stoplist words), called the "*B*-LIST", each of which appeared in at least 2 migraine titles and 2 magnesium titles.

Application of the stoplist to phrases merits further comment. The appropriateness of the stoplist for many words depends on the context in which they are used. "Protein", for example, is too broad to be useful, and so is put on the stoplist. But individual proteins are not, and the names of some of these include the word "protein", such as "amyloid beta-protein" or "protein kinase C". We therefore divide the stoplist into two parts, a short list of prepositions and other connectives and a long (5000-word) list. We apply only the short stoplist to each word in a phrase. Only if every remaining word in the phrase is on the long stoplist is the phrase dropped (Fig. 4).

The *B*-LIST is next edited by removing redundancies and nonuseful terms, resulting, for the migraine/magnesium case, in cutting the list from 200 to 150 terms. The editing process in general may include adding, deleting, removing redundancies, or revising entries in order to compensate for certain limitations in the mechanized rules.

After the editing has been completed, the computer produces a display or printed output that shows titles within title file *A* that share *B*-terms with titles in file *C*,

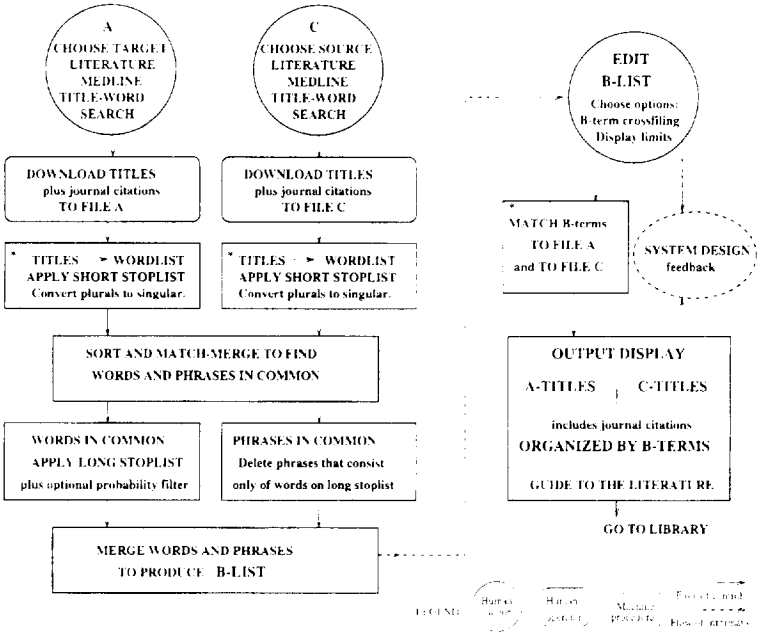

Fig. 4. A schematic flowchart of *Procedure* II. The procedure begins with a downloading of titles from literatures *A* and *C*, and proceeds to the lower right where an output display is produced as a heuristic aid for the human user of the system. The display is organized to facilitate comparison of *A*-titles with *C*-titles for each *B*-term that they have in common, and serves as a guide to the literature.

organized alphabetically by the 150 *B*-terms. A similar display is created for title file *C*, thus facilitating a comparison of titles *A* (magnesium) with titles *C* (migraine) that have one or more *B*-words or phrases in common. Fig. 5 shows the general format of the display and some of the more interesting titles selected for their suggestiveness of complementarity. This title display is the principal product of ARROWSMITH. In our example it contained altogether about 920 migraine titles and 1350 magnesium titles.

In general, portions of the title display output that appear promising on the basis of expert judgment then become a heuristic point of departure for reviewing literatures *A* and *C* in order to identify complementary statements or arguments (for example, *AB–BC* relationships) that lead to novel testable hypotheses.

Fig. 5. Selected entries from a printed display of the *Procedure* II output for two sets of titles that contain the word "magnesium" (*A*—left column) or "migraine" (*C*—right column), respectively. The headings shown are *B*-terms (in alphabetic sequence). Under each *B*-term are titles containing that term in the two columns of the display. This arrangement facilitates comparing "magnesium" titles with "migraine" titles for each *B*-term that they have in common. Shown here are examples of *B*-terms in which biologically meaningful relationships $A \Rightarrow B$ and $B \Rightarrow C$ are suggested by the titles; citations to the literature are also identified. Although, for each *B*-term, only a few titles are shown, in general there may be as many as several hundred. To avoid excessive output, a limit to the number of titles displayed for any one *B*-term can be set interactively by the user.

(a)

MAGNESIUM - A

CALCIUM CHANNEL
87272894
Mg2+-Ca2+ interaction in contractility of vascular smooth muscle: Mg2+ versus organic calcium channel blockers on myogenic tone and agonist-induced responsiveness of blood vessels. [Review] Can J Physiol Pharmacol. 65(4):729-45, 1987
87010519
Blockade of current through single calcium channels by Cd2+, Mg2+, and Ca2+. J Gen Physiol. 88(3):321-47, 1986
86256906
Microcirculatory actions and uses of naturally-occurring (magnesium) and novel synthetic calcium channel blockers. [Review] Microcirc Endothelium Lymphatics. 1(2):185-220, 1984

EPILEPSY
88053486
Neurological consequences of magnesium deficiency: correlations with epilepsy. Clin Exp Pharmacol Physiol. 14(5):361-70, 1987
79176542
Preliminary report on the magnesium deficient rat as a model of epilepsy. Lab Anim Sci. 28(6):680-5, 1978

EPILEPTIC
80020041
Epileptic-type convulsions and magnesium deficiency. Aviat Space Environ Med. 50(7):734-5, 1979
78106635
Effect of magnesium on epileptic foci. Epilepsia. 19(1):81-91, 1978
68011368
A case of the epileptic form of latent tetany due to magnesium deficit. Electroencephalogr Clin Neurophysiol. 23(4):388-9, 1967

HISTAMINE
87275485
Specific change of histamine metabolism in acute magnesium-deficient young rats. Drug Nutr Interact. 5(2):89-96, 1987.
87239284
Reduction of histamine-induced bronchoconstriction by magnesium in asthmatic subjects. Allergy. 42(3):186-8, 1987
86287572
Influence of magnesium on norepinephrine-and histamine-induced contractions of pulmonary vascular smooth muscle. Pharmacology. 33(1):27-33, 1986.
85261019
Effect of parenteral magnesium on pulmonary function, plasma cAmp, and histamine in bronchial asthma. J Asthma. 22(1):3-11, 1985

ISCHEMIA
84275376
Mild hypothermia and Mg++ protect against irreversible damage during CNS ischemia. Stroke. 15(4):695-98, 1984
85097212
Pharmacologic inhibition of cerebral vasospasm in ischemia, hallucinogen ingestion, and hypomagnesemia: barbiturates, calcium antagonists, and magnesium. Am J Emerg Med. 1(2):180-90, 1983

ISCHEMIC
87223584
The ionic basis of the anti-ischemic and anti-arrhythmic properties of magnesium in the heart. [Review] J Am Coll Nutr. 6(1):27-33, 1987
82156855
Relation of myocardial magnesium deficiency to sudden death in ischemic heart disease. Am Heart J 103(3):449-50, 1982
80147435
Magnesium deficiency produces spasms of coronary arteries: relationship to etiology of sudden death ischemic heart disease. Science. 208(4440):198-200, 1980

MIGRAINE - C

CALCIUM CHANNEL
86238440
Migraine treatment with calcium channel blockers. [Review] Acta Pharmacol Toxicol (Copenh). 58 Suppl 2:161-7, 1986.
85118978
Calcium-channel blockers in the treatment of migraine. [Review] Am J Cardiol. 55(3):139B-143B, 1985
84060905
The pharmacology of calcium channel antagonists: a novel class of anti-migraine agents?. Headache. 23(6):278-83, 1983
83212498
Flunarizine, a calcium channel blocker: a new prophylactic durg in migraine. [Review] Headache. 23(2):70-4, 1983

EPILEPSY
89064703
Migraine-epilepsy relationships. [Review] Epilepsy Res. 1(4):213-26, 1987
87277624
Is there a common pharmacological link between migraine and epilepsy? [Review] Funct Neurol. 1(4):515-20, 1986
69225134
The relation of migraine and epilepsy. Brain. 92(2):285-300, 1969.

EPILEPTIC
82064930
Association patterns between epileptic and migraine attacks. Acta Neurol (Napoli). 3(4):587-98, 1981
81031202
Basilar migraine? Seizures, and severe epileptic EEG abnormalities. Neurology. 30(10):1122-5, 1980

HISTAMINE
85298056
Role of histamine in the pathogenesis of migraine. [Review] Postepy Hig Med Dosw. 39(1):12-22, 1985
78047615
Histamine metabolism in cluster headache and migraine. Catabolism of 14C histamine. J Neurol. 216(2):105-17, 1977
67011667
Preventive treatment of migraine and histamine headache. Tidsskr Nor Laegeforen. 86(5):322-3, 1966 [Review]
66064937
The histamine-glucocorticoid relationship and its implications for migraine headache. Headache. 5(4):111-5, 1966

ISCHEMIA
87127625
Ischemia may be the primary cause of the neurologic deficits in classic migraine. Arch Neurol. 44(2):156-61, 1987
82110461
Impairment of cerebral serotonin and energy metabolism during ischemia: relevance to migraine. Adv Neurol. 33:35-40, 1982.
83081984
Migraine as a model of neurogenic ischemia [editorial] Headache. 22(6):287-8, 1982

ISCHEMIC
87156114
The ischemic hypotheses of migraine. Arch Neurol. 44(3):321-2, 1987
85086750
Migrainous ischemic optic neuropathy. Neurol 35(1):112-4, 1985
83191503
Migraine, a risk factor for ischemic cerebral stroke. Med Welt. 34(8):233-4, 1983

Fig. 5(a).

(b)

MAGNESIUM - A

PROSTAGLANDIN

80235079
Magnesium ions control prostaglandin reactivity of venous smooth muscle from spontaneously hypertensive rats.
Prostaglandins Med. 4(4):255-61, 1980

74289399
Effects of magnesium ion and oxytocin inhibitors on the uterotonic activity of oxytocin and prostaglandins E2 and F2alpha.
J Pharmacol Exp Ther. 190(1):77-87, 1974

REACTIVITY

86229235
Vasomotor effects of magnesium: a comparison with nifedipine and verapamil of in vitro reactivity in feline cerebral and peripheral arteries. Magnesium. 5(2):66-75, 1986.

80235079
Magnesium ions control prostaglandin reactivity of venous smooth muscle from spontaneously hypertensive rats.
Prostaglandins Med. 4(4):255-61, 1980

78124525
Magnesium and vascular tone and reactivity.
Blood Vessels. 15(1-3):5-16, 1978.

SEROTONIN

80248760
Aromatic amines (serotonin and histamine) and magnesium deficiency in the rat. Int J Vitam Nutr Res. 50(2):185-92, 1980.

79224036
Effects of cerebral intraventricular magnesium injections and a low magnesium diet on nonspecific excitability level, audiogenic seizure susceptibility and serotonin.
Pharm Biochem Behav. 10(4):487-91, 1979

76171101
Cerebral arterial spasm. Part 4: in vitro effects of temperature, serotonin analogues, large nonphysiological concentrations of serotonin, and extracellular calcium and magnesium on serotonin-induced contractions of the canine basilar artery.
J Neurosurg. 44(5):585-93, 1976

SPREADING DEPRESSION

87168648
Low extracellular magnesium induces epileptiform activity and spreading depression in rat hippocampal slices.
J Neurophysiol. 57(3):869-88, 1987

85056949
The nature of the chick's magnesium-sensitive retinal spreading depression. J Neurobiol. 15(5):333-43, 1984

VASOSPASM

85097212
Pharmacologic inhibition of cerebral vasospasm in ischemia, hallucinogen ingestion, and hypomagnesemia: barbiturates, calcium antagonists, and magnesium. Am J Emerg Med. 1(2):180-90, 1983

81075542
Withdrawal of magnesium causes vasospasm while elevated magnesium produces relaxation of tone in cerebral arteries.
Neurosci Lett. 20(3):323-7, 1980

MIGRAINE - C

PROSTAGLANDIN

81061951
Use of a prostaglandin inhibitor in migraine crisis. Study of 40 cases. Arq Neuropsiquiatr. 38(2):140-3, 1980

77255822
Ophthalmoplegic migraine: amelioration by Flufenamic acid, a prostaglandin inhibitor. Ophthalmologica. 175(3):148-52, 1977

76174742
Migraine attacks. Alleviation by an inhibitor of prostaglandin synthesis and action. Neurology. 26(5):447-50, 1976

REACTIVITY

83009859
Extracranial and cardiovascular reactivity in migrainous subjects.
J Psychosom Res. 26(3):317-31, 1982.

82259333
Extracranial vascular reactivity in migraine and tension headache.
Cephalalgia. 1(3):149-55, 1981

79238926
Abnormal cerebrovascular reactivity in patients with migraine and cluster headache. Headache. 19(5):257-66, 1979

72069533
Reactivity of the intra- and extracranial vessels to serotonin and its relation to migraine. Gac Med Mex. 100(12):1297-308, 1970

SEROTONIN

82110461
Impairment of cerebral serotonin and energy metabolism during ischemia: relevance to migraine. Adv Neurol. 33:35-40, 1982.

85099289
Platelet aggregability, disaggregability and serotonin uptake in migraine. Cephalalgia. 4(4):221-5, 1984

82110462
Serotonin precursors in migraine prophylaxis.
Adv Neurol. 33:357-63, 1982.

82110455
The evolution of thinking about the role and site of action of serotonin in migraine. Adv Neurol. 33:31-3, 1982.

SPREADING DEPRESSION

88129711
Cerebral blood flow in migraine and cortical spreading depression. Acta Neurol Scand Suppl 113:1-40, 1987 [Review]

88165004
Characteristics of spreading depression and of its propagation. Their possible role in migraine. Cephalalgia 7 Supp 6 65-8, 1987

86013123
Is migraine explained by Leao's spreading depression?.
Lancet. 2(8458):763-6, 1985

85197227
Experiments on spreading depression in relation to migraine and neurosurgery. An Acad Bras Cienc. 56(4):423-30, 1984 Dec.

VASOSPASM

85029778
Bilateral cervical carotid and intracranial vasospasm causing cerebral ischemia in a migrainous patient: a case of "diplegic migraine". Headache. 24(5):245-8, 1984

84113760
Isolated benign cerebral vasculitis or migrainous vasospasm?.
J Neurol Neurosurg Psychiatry. 47(1):73-6, 1984

79193732
Vasospasm and vascular headaches: selective vasoconstriction in the carotid vascular system measured by the Doppler ophthalmic method in migraineurs. Headache. 19(4):200-3, 1979

67256447
Cerebral vasospasm in migraine. J Lancet. 87(8):283-6, 1967

Fig. 5(b).

MIGRAINE-MAGNESIUM B-LIST (selected)

BC	AB	B-term	
5	3	amine	
3	2	anticonvulsant	
5	2	calcium antagonist	*
10	2	calcium channel	*
4	1	calcium entry	*
5	3	catecholamine	
5	8	diabetes	
3	3	dopamine	
14	2	epilepsy	*
5	6	epileptic	*
11	11	hemodynamic	
14	13	histamine	
11	3	ht	*
15	4	hydroxytryptamine	*
3	11	hypertension	
3	2	hypoxia	*
6	3	immunoglobulin	
3	7	inflammatory	*
2	0	ischaemia	
12	3	ischemia	
6	1	ischemic	
9	8	muscle contraction	
5	4	olfactory	
14	5	oral contraceptive	
10	3	paroxysmal	
14	5	platelet aggregation	*
4	2	progesterone	
14	4	prolactin	
10	3	prolapse	
12	5	prostaglandin	*
8	3	reactivity	*
16	7	relaxation	
10	7	reserpine	
8	14	seizure	*
11	5	serotonin	*
4	4	spasm	*
5	2	spreading depression	*
7	5	stress	*
6	7	tryptophan	
4	5	vasospasm	*
6	4	verapamil	*
322	192	41 TOTALS	

RAYNAUD-FISHOIL B-LIST

BC	AB	B-term	
1	1	angina	
2	2	arthritis	
2	5	blood pressure	
10	5	blood viscosity	*
6	7	calcium	
12	1	capillary	
2	1	collagen	
4	2	deformability	*
1	5	diabetic	
3	1	fibrinolytic	
1	1	hemolytic uremic syndrome	
9	2	hypertension	
1	4	hypertensive	
1	1	iga	
3	3	infarction	
1	3	inhibition platelet	*
1	5	ischemic	
8	2	lupus	
1	1	mediterranean	
2	1	pgi2	
2	13	platelet aggregation	*
3	14	platelet function	*
1	1	polymorphonuclear	
10	9	prostacyclin	
10	25	prostaglandin	*
2	1	prostaglandin i2	
1	1	reactivity	*
1	1	serotonin	*
1	2	thrombotic	
6	11	thromboxane	*
1	2	thyroid	

Fig. 6. Selected entries from the *B*-LIST for magnesium and migraine, and for fish oil and Raynaud's disease, produced by *Procedure* II (Fig. 4). The two numbers shown in the two columns to the left of each word list represent the number of articles within the *BC* and *AB* intersections, respectively, as illustrated in Fig. 1. The asterisks mark entries identified in the original studies [29,31].

A study of the output display for magnesium and migraine revealed about 40 *B*-terms (listed in Fig. 6) that merit further investigation because they appear, from some of the titles with which they are associated, to be related to both magnesium and migraine in a way that suggests a physiological linkage.

3. Evaluating ARROWSMITH

In this paper we try to make clear how and why we found ARROWSMITH useful in a search for complementary literatures. We have presented a step by step process and detailed examples of output to permit others to form a reasonable judgment concerning usefulness.

Prior to designing ARROWSMITH, we had available three completed and published analyses of complementary noninteractive literatures [29,31,33]. These studies provide an opportunity to determine whether ARROWSMITH can at least be helpful in rediscovering complementary structures already known by the user to exist. Such an outcome cannot be taken for granted without being put to a test — principally because it is not otherwise obvious that article titles alone contain enough information to put the user on the track of complementary literatures.

For the migraine case study exemplified in the preceding section, we have noted that "magnesium" was first on the ranked list of terms produced automatically as the output of *Procedure* I. The title word co-occurrence data that led to this ranking was not generated at the time of the original study; magnesium was chosen on quite different grounds at that time. Not only did magnesium rank high on the list of *A*-candidates, but within the *B*-list generated by *Procedure* II (Fig. 6), one can recognize ten of the eleven connections identified originally in the 1988 migraine/magnesium study (the eleventh did not co-occur with "magnesium" as a title word in MEDLINE). In addition, a substantial number of new and plausible candidates for *B*-terms also appeared on the list. The 19 words and phrases marked with an asterisk in Fig. 6 correspond, with redundancies removed, to the 10 terms discussed in the earlier study [31]. Thus, using *Procedure* II, almost all of the originally reported *B*-term connections were produced automatically, and in a matter of hours, replacing weeks long literature searching and exploration that was previously required.

Up to this point, production of the ranked *A*-list and the *B*-list are automatic, provided the user chooses the same search restriction categories and strategy (based on terms related to dietary intake) as we exemplified. Users do have the option of changing this feature of the exploratory process however. In the next stage, the production of the output title display (of which a small part is shown in Fig. 5) is also automatic, for it is fully defined by the *B*-list and the original set of downloaded titles. Fig. 5 itself was not automatically produced; the 67 titles shown were selected from the complete display of about 2270 titles. This selection process is of course sensitive to what the user is looking for, and other users might not see the same, or as many, linked titles in the output display as we found. However, we invite readers to study Fig. 5 closely and try to form a judgment of the effectiveness of these word linked titles in suggesting a possible link between magnesium deficiency and migraine. Successful use of the complete title display does not depend on recognizing or selecting all of the titles we selected, but only on seeing enough linked titles to impel further searching and analysis. Thus, the method is probably quite robust with respect to individual variability among users. Accordingly, we believe that the success of this retrospective application of *Procedure* I in discovering an already known structure provides a reasonable basis for judging how the system would have performed had the structure not been known in advance. That is, the

opportunity for favorably distorting the outcome because the users knew what to look for is relatively limited.

The application of *Procedure* I to the pre-1986 literature on Raynaud's disease [29] was similarly successful, except that here the context revealed in FILE *U* played a more central role. "Oil", one of the high ranking but fairly broad words on the *A*-list (Fig. 3) occurred (within FILE *U*) repeatedly in the same, and much narrower, context defined by "dietary fish oil" (or equivalent terms). Fig. 6 shows a *B*-list for the Raynaud's disease analysis. We produced substantially the same outcome as in the original analysis reported in 1986 [28,29]. In addition, the main findings of that analysis have been largely replicated, as well as extended, independently by other researchers [8].

The success of ARROWSMITH in the above two case studies was not matched in the third study, that of somatomedin C and arginine [33]. This attempt failed because there were no paired titles that were suggestive of complementary arguments within the corresponding literatures. This failure to re-discover an already known structure underscores an important point. We cannot and do not claim that the procedures we have developed will always be successful, even in the hands of a prescient user. Some of the many reasons why our procedures, as presently designed, might fail to reveal complementary literatures that do exist are discussed in Section 5.2.

The next section of this paper can also be seen as an evaluation of ARROWSMITH (*Procedure* II). But in this case, the test is focused on producing novel results rather than on finding structures that we previously had discovered using more conventional literature search methods.

4. Examples that use Procedure II independently of Procedure I

Procedure II will be useful independently of *Procedure* I if the user already has formed hypotheses concerning plausible *A* literatures. For example, a biomedical researcher may have reason to think that a particular *A*-term is related to *C* without necessarily knowing the intermediate pathways or mechanisms by which *A* and *C* are linked. *Procedure* II can be regarded as a "higher order MEDLINE search" (because it finds 2nd order indirect title word connections). The following three examples fit this pattern. The first two examples, on Alzheimer's disease (AD), show how previously unknown connections can be discovered even between two literatures already known to interact in other respects.

4.1. Case example 1: indomethacin and Alzheimer's disease (AD)

The literatures corresponding to indomethacin (*A*) and Alzheimer's disease (*C*) intersect and are interactive. Indeed, indomethacin appears to have a protective effect against AD, as evidenced by epidemiologic and clinical data [22]. Nevertheless indomethacin, an inhibitor of prostaglandin synthesis, affects many organ systems and its effects have been reported in the literatures of many diverse specialties. Therefore it is not an easy task to find all known effects of indomethacin which are likely to be relevant to patients with AD. *Procedure* II offers a useful aid to solving this problem by

displaying numerous plausible indirect (A–B–C) title word linkages between the two literatures on indomethacin and AD. In this case ARROWSMITH found 103 B-terms, of which 5 referred to substances or physiologic processes affected by indomethacin in a way that, as indicated separately in the AD literature, might possibly ameliorate AD [22]. In addition, we found a possible adverse effect: indomethacin had a well-documented anti-cholinergic action, and separately it was clear from the AD literature that one would expect anti-cholinergic action to affect AD patients adversely by exacerbating cognitive dysfunction. The possibility that indomethacin might have adverse side effects in AD patients apparently had not been considered explicitly in any scientific publication and so we brought this to the attention of neuroscientists [22].

4.2. Case example 2: estrogen and Alzheimer's disease

Even though it has become relatively well established that estrogen replacement therapy for postmenopausal women results in a lower incidence of AD, the mechanism of such beneficial effects is not understood. We used *Procedure* II to develop a list of potential B-linkages between estrogen and AD; there were several hundred B-terms that appeared in at least 2 titles in each literature. About 8 of these terms (calbindin D28K, cathepsin D and other proteases, superoxide dismutase, antioxidants, apolipoprotein E, glutamate, cytochrome C oxidase subunit III) refer to substances that are known to be modulated by estrogen and are separately known to be implicated in AD, but which have not been investigated as estrogen targets in the context of AD. In the process of investigating the literature identified by the output title display, we encountered intriguing reports indicating that estrogen has antioxidant activity. There is also an extensive AD literature on free radicals (groups of atoms, usually highly active, that have one or more unpaired electrons) which are thought to play a role in the development of AD. Because antioxidants would tend to counteract such effects, it is remarkable that there was no evidence in the database accessible literature that AD researchers were aware of this possible mechanism for estrogen's beneficial effects [23]. After our paper was submitted for publication [23], one study did appear which mentioned this link.

4.3. Case example 3: phospholipases and sleep

The hypothesis that phospholipases may regulate sleep has been raised implicitly (phospholipases may regulate prostaglandin synthesis and several prostaglandins are endogenous sleep modulators), but this idea has not been investigated experimentally nor discussed explicitly in the biomedical literature. We employed *Procedure* II in three searches using "sleep" as the C literature and "phospholipase A_2", "phospholipase C", or "phospholipase D" as A literatures. (In this case we found that most B-terms did not represent substances formed by phospholipase activity, but rather substances that stimulate or inhibit phospholipases.) We found a number of cytokines and neuromodulators among the B-terms, notably the major sleep promoting substances interleukin 1β, tumor necrosis factor, and endotoxin/lipopolysaccharide. These substances stimulate phospholipase A_2, C, and D activities in various systems, and this stimulation is thought to be required for at least some of their biological activities. Thus, *Procedure* II

generated a list of promising agents whose effects on sleep may involve phospholipases, and we suggest testing whether specific phospholipase inhibitors inhibit the sleep promoting effects of these agents [24].

5. Discussion

5.1. Intelligence, implanted structures, and the role of heuristics

ARROWSMITH provides three stages of output information structures that serve as heuristic guides to complementary scientific arguments within the literature: (i) the list of *A*-candidates (Fig. 3); (ii) the *B*-LIST; and (iii) the title display organized by *B*-terms (Fig. 5). The process is heuristic in that human choices at each stage not only are assisted by the displayed structures [40], but these choices in turn influence the outcome of each later stage. Each cycle through the three stages provides valuable information for a retry, either with a new choice from the same *A*-candidate list, or possibly with a new *A*-list derived from a different restriction category.

ARROWSMITH possesses no algorithm for recognizing "interesting" relationships such as transitivity or complementarity, but it *seems* to have such a capability. The density of interesting title relationships revealed by the stage (iii) display appears to be remarkably greater than the density of interesting relationships among titles selected randomly even from two complementary literatures. The consequent impression of intelligence arises primarily from three sources: the stoplist, the search strategy restriction category, and the organization of the output display. These structures are human created, then supplied to the computer, and can be thought of as implanted intelligence. ARROWSMITH filters and organizes title data based on these implanted structures.

5.2. Limitations, improvements and future directions

The use of title words as a basis for detecting complementary literatures is both a strength and a weakness of ARROWSMITH. The advantage of titles lies in the fact that they provide a constrained context within which the linkages between *A*- and *B*-terms, and between *B*- and *C*-terms, tend to be biologically meaningful and easily perceived by the viewer, thus facilitating the recognition of potential complementarity. On the other hand, ARROWSMITH at present does not attempt automatic linkage detection in abstracts or subject headings. Other investigators have shown that such information can be exploited. Using full MEDLINE records (1983–1985 only), no restriction categories, and several statistical criteria, Gordon and Lindsay have confirmed that fish oil is a high ranking *A*-candidate in the *A*–*B*–*C* model for Raynaud's disease [8].

Each of ARROWSMITH's output heuristic information structures serves also as a source of feedback that is potentially useful for improving the system, evaluating it, and perhaps for stimulating ideas that can lead to better models and theories. Within the framework of the present design, numerous incremental improvements appear to be possible through augmenting the stoplist, improving phrase recognition by allowing for more variation in word morphology and word position, organizing and grouping

B-terms with the help of subject headings, extending the *A*-candidate list to include phrases, and developing a synonym recognition capability.

The most practical and straightforward approach to improving ARROWSMITH consists in juxtaposing pairs of titles not only on the basis of words they have in common but in addition by common subject headings in the two respective MEDLINE records. This kind of linkage supplies additional context for paired titles that may be meaningful to the expert observer. Moreover, linking titles by common subject headings may also reduce the number of near synonymous or equivalent structures that are missed by word matching.

On a longer range basis, we envision applying ARROWSMITH to databases other than MEDLINE and particularly to records that may be derived from different databases. For such applications, the Metathesaurus® of the Unified Medical Language System® (UMLS) offers a potentially valuable resource for developing a hierarchy/synonym recognition capability. The 1996 Metathesaurus is a synthesis of 38 different source vocabularies (one of which is *Medical Subject Headings* (*MeSH*)). It contains over 250,000 concepts named by 590,000 biomedical terms, and reflects all names, meanings, and hierarchical and inter-term relationships that are present in the original source vocabularies. In addition, it establishes new relationships between terms from different source vocabularies [20].

Significant advances beyond the capabilities of existing database search systems might follow from fundamental improvements in the indexing of relationships. Existing index languages and subject heading systems (such as *MeSH*) reflect context independent language structures, particularly synonymous and hierarchical relationships. These systems are not designed or intended to represent adequately the relationships between different subject headings that are appropriate for a particular article to which the headings are applied (hence are context dependent) but not appropriate for all articles and contexts. For example, an article discussing arteriosclerosis as a possible cause of hypertension would be indexed under the subject headings "arteriosclerosis" and "hypertension"; there is no provision for linking the two terms together, in the process of indexing a given article, to show that a cause effect relationship between them is discussed in that article. The hierarchical relationship between "arteriosclerosis" and "coronary disease", on the other hand, is valid in all contexts and hence is built into the MeSH structure itself. The requirement for improved indexing of context dependent relationships calls attention to these relationships themselves as entities that merit more systematic study [9,10], and has led to important research on interactive frame based or knowledge based indexing [9,10,14,15].

5.3. Comments on methodology

Any two literatures or sets of titles that share a common language will have many words in common irrespective of whether the two literatures have substantial or interesting scientific linkage. Hence we must address the question of whether ARROWSMITH, in manipulating title words, really helps reveal linkage of scientific interest. The *A*-list consists of words that co-occur in titles with other words (the *B*-terms) that co-occur in titles with *C*-terms. The list is subjected to three filtering processes—an

extensive stoplist, a probabilistic cutoff that retains only words strongly correlated (A with B and B with C) and a category restriction. Our results at least suggest that the more important and better established A–B and B–C biological connections tend to pass through the filters. Even though the problem of finding nuggets of value among a large number of uninteresting connections persists, the above process of enrichment appears sufficient to permit the more interesting or important connections to be spotted easily by an expert. It is also our strong impression that the high ranking biologically plausible A-terms lead to a more extensive and much richer B-LIST and title display in *Procedure* II than do words of equal *a priori* plausibility but which are of low rank or not on the A-list at all.

We have defined the complementarity of two literatures based on the scientific arguments and discourse within those literatures, not on just shared language use. To examine and illustrate this distinction further, we developed a new A-candidate list (Fig. 3), for the literature on Alzheimer's disease (AD), using a restriction category based on toxins or poisons [36], and analyzed the underlying biologic connections associated with one high rank candidate, lead.

The appearance of the poisonous metal "lead" among the top ranked words on the A-list for AD invites attention. Lead is well known as an environmental hazard with neurotoxic effects. Among the various high ranking terms on the A-list that would seem worth investigating (including ozone, ethanol, TNF (tumor necrosis factor), IL (interleukin), cadmium) lead is a conspicuous and biologically plausible choice. Even though all A-words participate in A–B–C *word* linkages, these linkages are not equally plausible for AD. For example, the A–B linkages for ozone tend to affect the lungs but those for lead tend to affect the brain. Thus lead is judged to be a more promising candidate.

But it should not be surprising that the quite large literatures on lead and AD share a great deal of common language, including B-terms in titles, irrespective of whether lead and AD have meaningful biologic interaction. To pursue this question further, we examined not only the title linkages, but the underlying literatures as well. Specifically, 5 groups of effects could be identified in the separate literatures on lead and AD: (i) lead treatment or lead exposure alters cholinergic function in a number of systems; deficient cholinergic function is a hallmark of AD. Lead exposure was reported to cause a decrease in choline acetyltransferase in the septum, a brain area known to be particularly deficient in this enzyme in AD patients. (ii) Lead has effects on calcium dependent signal transduction pathways involving adenylate cyclase, calmodulin, and kinase; all three signal transduction pathways are reported to be abnormal in AD and are thought to participate in its pathogenesis. (iii) It is well established that lead affects brain vasculature and glial cells, and alters the blood brain barrier; these are also reported to be abnormal in AD. (iv) Lead exposure is associated with oxidative stress that generates free radicals (see Section 4.2); many reports have indicated that excess free radical formation occurs in AD. (v) Lead exposure interferes with long term potentiation in the hippocampus, a cellular model of short term memory; the latter is known to be abnormal in AD.

The connections that were found appear to go far beyond the matter of "just" shared language, and indeed suggest that the cellular effects of lead exposure entail functional

deficits that also accompany AD. We cannot conclude on that basis that lead exposure is necessarily a significant risk factor in AD, for this inference would require direct evaluation of lead in living systems. But the foregoing argument does tend to support our claim that the *A*-list for Alzheimer's disease identified agents having meaningful *A–B–C* linkages with the source literature.

6. Application to other fields and disciplines

The question has often been raised of whether our techniques can be applied to fields other than biology and medicine. In principle it is possible, but in practice the answer may depend on the accuracy and specificity of article titles; in biology and medicine, titles tend to be highly specific and informative. The same techniques could be applied of course to abstracts, but we haven't tried it and do not know whether the advantages of establishing more connections will be outweighed by the disadvantages of more "noise" in the system. Moreover, all of the examples we have worked on so far have been in the subset of biomedical literature that is relevant to mammalian physiology. In this corner of the "real world", everything is richly interconnected. Thus each publication about some aspect of that world is likely to create an intricate web of implicit and unintended connections within the corresponding world of recorded knowledge. Fields that are not as intimately interconnected may be less rewarding for our methods. In the field of chemistry, our *A–B–C* model could refer to a chemical reaction pathway and would have in that case interesting similarities and contrasts to work in AI on the automatic generation of reaction pathways [38].

7. Conclusions and implications

Bringing together two complementary but noninteractive literatures from different specialized areas of research can reveal undiscovered public knowledge — unnoticed implicit relationships not apparent in the two literatures considered separately. This paper has described an experimental "discovery support system" [8] called ARROWSMITH which embodies a replicable database search procedure and related software that produces heuristic aids to finding complementary literatures and to deriving novel scientific hypotheses. Applying ARROWSMITH has led to testable hypotheses, several of which have been experimentally corroborated, that demonstrate the effectiveness and value of the system.

ARROWSMITH, operating in an environment characterized by the unruly problems of natural language text and the immensity of the scientific record, is a practical system that seeks immediate results in furthering the aims of biomedical research. At the same time, it is a research tool for studying undesigned but human created structures in the literature of science. Each use of the system creates numerous examples of word linked titles suggestive of complementarity that are of potential value in examining the logic of scientific discourse, in new approaches to the indexing of relationships [9,10,14,15], and as a source of ideas for modelling the process of discovery. Literature based discoveries

may also open new lines of investigation in the growth of knowledge and the drives toward specialization and fragmentation in science [18,19,39,41,42]. We invite inquiries from those who might wish to try using the system on problems of their own.

References

[1] A. Bookstein and D.R. Swanson, Probabilistic models for automatic indexing, *J. Am. Soc. Inf. Sci.* **25** (1974) 312–318.
[2] B.B. Chang, R. DiGiacomo, J. Kremer, C. Kay and D.M. Shah, Effects of fish oil fatty acid ingestion in patients with Raynaud's syndrome, *Surgical Forum* **39** (1988) 324–326.
[3] Z. Chen, Let documents talk to each other: a computer model for connection of short documents, *J. Doc.* **49** (1993) 44–54.
[4] R. Davies, The creation of new knowledge by information retrieval and classification, *J. Doc.* **45** (1989) 273–301.
[5] R.A. DiGiacomo, J.M. Kremer and D.M. Shah, Fish-oil dietary supplementation in patients with Raynaud's phenomenon: a double-blind, controlled, prospective study, *Am. J. Med.* **86** (1989) 158–164.
[6] A.R. Gardner-Medwin, Possible roles of vertebrate neuroglia in potassium dynamics, spreading depression and migraine, *J. Exp. Biol.* **95** (1981) 111–127.
[7] E. Garfield, Linking literatures: an intriguing use of the Citation Index, *Current Contents* **21** (1994) 3–5.
[8] M.D. Gordon and R.K. Lindsay, Toward discovery support systems: a replication, re-examination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil, *J. Am. Soc. Inf. Sci.* **47** (1996) 116–128.
[9] R. Green, Syntagmatic relationships in index languages: a reassessment, *Libr. Q.* **65** (1995) 365–385.
[10] R. Green, Topical relevance relationships. I. Why topic matching fails, *J. Am. Soc. Inf. Sci.* **46** (1995) 646–653.
[11] S.P. Harter, Scientific inquiry: a model for online searching, *J. Am. Soc. Inf. Sci.* **35** (1984) 110–117.
[12] S.P. Harter, *Online Information Retrieval Concepts, Principles, and Techniques* (Academic Press, New York, 1986).
[13] S.P. Harter and A.R. Peters, Heuristics for online information retrieval: a typology and preliminary listing, *Online Rev.* **9** (1985) 407–424.
[14] S.M. Humphrey, Indexing biomedical documents—from thesaural to knowledge-based retrieval systems, *Artif. Intell. Med.* **4** (1992) 343–371.
[15] S.M. Humphrey and N.E. Miller, Knowledge-based indexing of the medical literature: the indexing aid project, *J. Am. Soc. Inf. Sci.* **38** (1987) 184–196.
[16] D. Kulkarni and H.A. Simon, The processes of scientific discovery: the strategy of experimentation, *Cognit. Sci.* **12** (1988) 139–175.
[17] P. Langley, H.A. Simon, G.L. Bradshaw and J.M. Zytkow, *Scientific Discovery, Computational Explorations of the Creative Process* (MIT Press, Cambridge, MA, 1987) Chapter 10, 302–307.
[18] J. Lederberg, Communication as the root of scientific progress, in: E. Garfield, ed., *Essays of an Information Scientist* **15** (ISI, Philadelphia, PA, 1993) 208–214.
[19] R.K. Merton, *The Sociology of Science* (University of Chicago Press, Chicago, IL, 1973) 371–376.
[20] National Library of Medicine (1996): Web page. http://www.nlm.gov/publications/factsheets/umls_metathesaurus.html or ftp from nlmpubs.nlm.nih.gov/umls/meta.txt.
[21] N.R. Smalheiser and D.R. Swanson, Assessing a gap in the biomedical literature: magnesium deficiency and neurologic disease, *Neurosci. Res. Commun.* **15** (1994) 1–9.
[22] N.R. Smalheiser and D.R. Swanson, Indomethacin and Alzheimer's disease, *Neurol.* **46** (1996) 583.
[23] N.R. Smalheiser and D.R. Swanson, Linking estrogen to Alzheimer's disease: an informatics approach, *Neurol.* **47** (1996) 809–810.
[24] N.R. Smalheiser and D.R. Swanson, Use of a novel database search strategy to aid in assessing hypotheses: do phospholipases regulate sleep? (in preparation).

[25] H.G. Small, Co-citation in the scientific literature: a new measure of the relationship between two documents, *J. Am. Soc. Inf. Sci.* **24** (1973) 265–269.
[26] M.C. Smeets, C.B. Vernooy, J.H. Souverijn and M.D. Ferrari, Intracellular and plasma magnesium in familial hemiplegic migraine and migraine with and without aura, *Cephalalgia* **14** (1994) 29–32.
[27] M. Spasser, The enacted fate of undiscovered public knowledge, *J. Am. Soc. Inf. Sci.* (to appear).
[28] D.R. Swanson, Undiscovered public knowledge, *Libr. Q.* **56** (1986) 103–118.
[29] D.R. Swanson, Fish oil, Raynaud's syndrome, and undiscovered public knowledge, *Perspect. Biol. Med.* **30** (1986) 7–18.
[30] D.R. Swanson, Two medical literatures that are logically but not bibliographically connected, *J. Am. Soc. Inf. Sci.* **38** (1987) 228–233.
[31] D.R. Swanson, Migraine and magnesium: eleven neglected connections, *Perspect. Biol. Med.* **31** (1988) 526–557.
[32] D.R. Swanson, Online search for logically-related noninteractive medical literatures: a systematic trial-and-error strategy, *J. Am. Soc. Inf. Sci.* **40** (1989) 356–358.
[33] D.R. Swanson, Somatomedin C and arginine: implicit connections between mutually-isolated literatures, *Perspect. Biol. Med.* **33** (1990) 157–186.
[34] D.R. Swanson, Medical literature as a potential source of new knowledge, *Bull. Med. Libr. Assoc.* **78** (1990) 29–37.
[35] D.R. Swanson, The absence of co-citation as a clue to undiscovered causal connections, in: C.L. Borgman, ed., *Scholarly Communication and Bibliometrics* (Sage Publ., Newbury Park, CA, 1990) 129–137.
[36] D.R. Swanson, Complementary structures in disjoint science literatures, in: A. Bookstein, Y. Chiaramella, G. Salton and V.V. Raghavan, eds., *SIGIR '91: Proceedings of the Fourteenth Annual International ACM / SIGIR Conference on Research and Development in Information Retrieval*, Chicago (ACM Press, New York, 1991) 280–289.
[37] D.R. Swanson, Intervening in the life cycles of scientific knowledge, *Libr. Trends* **41** (1993) 606–631.
[38] R.E. Valdés-Pérez, Conjecturing hidden entities by means of simplicity and conservation laws: machine discovery in chemistry, *Artif. Intell.* **65** (1994) 247–280.
[39] R.D. Whitley, The fragmentation of the sciences: remarks on the decline of university disciplines as units of knowledge production and evaluation, *Commun. Cognit.* **12** (1979) 363–370.
[40] W.C. Wimsatt, Heuristics and the study of human behavior, in: D.W. Fiske and R. Schweder, eds., *Metatheory in Social Science, Pluralisms and Subjectivities* (University of Chicago Press, Chicago, IL, 1986) 293–314.
[41] J.M. Ziman, The proliferation of scientific literature: a natural process, *Science* **208** (4442) (1980) 369–371.
[42] H. Zuckerman and J. Lederberg, Postmature scientific discovery?, *Nature* **324** (6098) (1986) 629–631.