ELSEVIER

Computer Methods and Programs in Biomedicine 57 (1998) 149–153

Computer Methods and Programs in Biomedicine

Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses

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

[a] *Department of Psychiatry, University of Illinois, M/C 912, 1601 W. Taylor Street, Chicago, IL 60612, USA*
[b] *Division of the Humanities, University of Chicago, 1010 E. 59th Street, Chicago, IL 60637, USA*

Received 6 February 1998; accepted 24 February 1998

Abstract

Conventional computer searches of the biomedical literature (e.g. MEDLINE) allow investigators to retrieve much of the information that has already been published on a given topic. However, these searches are of limited utility at the frontier of scientific discovery, when one wishes to identify and assess new, untested scientific hypotheses, or to uncover biologically significant relations between two previously disparate fields of inquiry. We have designed a set of interactive software and database search strategies, collectively called ARROWSMITH, that facilitate the discovery of plausible hypotheses linking findings across specialties (Artif. Intell. 91 (1997) 183–203). In the simplest implementation of ARROWSMITH, the user begins with an experimental finding or hypothesis that two items A and C are related in some way. The titles of papers indexed in MEDLINE which contain the word 'A' (or synonyms) are downloaded into a file A, and similarly a file C is created. The software constructs a list of words and phrases B common to files A and C; automatic and manual editing are used to filter out uninteresting B-terms. For each B-term, the software generates an AB file of titles containing both 'A' and 'B', and a BC file of titles containing both 'B' and 'C'; these titles are juxtaposed to facilitate the user judging whether there is likely to be a biologically significant relation among A, B and C. ARROWSMITH has been employed to analyze research problems relating to oxidative stress, brain damage, Alzheimer's disease and schizophrenia. Applications of ARROWSMITH include: anticipating adverse drug reactions, identifying mechanisms by which agents modulate cellular or organismal responses, suggesting new therapeutic approaches, identifying possible risk factors for diseases, and identifying potential animal models for human conditions. A simplified experimental version of ARROWSMITH is now freely accessible on the World Wide Web (http://kiwi.uchicago.edu).

Keywords: Information retrieval; Scientific discovery; Drug discovery; Alzheimer's disease; Oxidative stress

* Corresponding author. Tel.: +1 312 4134581; fax: +1 312 4134544; e-mail: smalheiser@psych.uic.edu

0169-2607/98/$ - see front matter
PII S0169-2607(98)00033-9

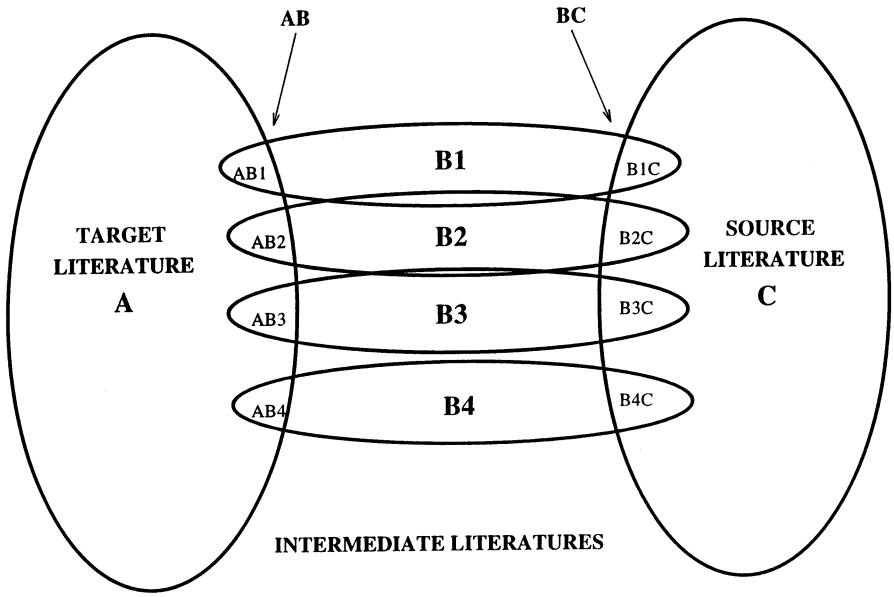

Fig. 1. A Venn diagram that represents sets of articles, or literatures, containing the words A and C in their titles. Sets A and C are linked through intermediate sets B_i ($i = 1, 2, 3\ldots$) which contain the word B_i in their titles and which overlap both A and C. By examining the articles in the pairs of intersections AB_i and B_iC, useful information may be inferred regarding possible biological linkages among A, B and C. (A and C are shown here as having no articles in common. When there is overlap between sets A and C, the articles in the direct intersection should first be identified and evaluated prior to carrying out an ARROWSMITH search.) Modified from [9] with permission.

1. Introduction

In the 1980s, one of us (D.S.) proposed a concept known as 'undiscovered public knowledge' [1], which may be illustrated with an example from the field of genetics: The cloning of new genes, which once was strictly a laboratory effort, has become greatly facilitated with the advent of the large genomics databases. Given a short stretch of nucleotide sequence, one can search the expressed sequence tag (EST) or other databases for homologous or overlapping sequences, and then for those which overlap with the second sequence, and so on. In a recent case, workers reconstructed the entire coding region of a hitherto-undiscovered gene solely from information residing within databases [2]. Each of the pieces of this gene were public knowledge, yet it still remained for these workers to identify them and link them to form a coherent whole which had not been noticed before (undiscovered).

Undiscovered public knowledge is not restricted to genomic databases, but exists in text-based databases such as MEDLINE as well [3–9]. Not only pieces of DNA sequences, but fragments of text that represent scientific knowledge can be identified and linked together (Fig. 1). For example, one experimental paper may report that A influences B; another paper, published in some other journal at some other time, may report that B influences C. The inference A influences C will represent undiscovered public knowledge if the statement has never been experimentally tested before, yet is confirmed once experimental tests are performed. Given the daunting number of possible linkages in the biomedical literature, the challenge is to find a practical method for identifying and linking items which helps the user to discern hypotheses that are intriguing, plausible, novel and testable. Alternatively, one may begin with the statement that A influences C (which may represent either an experimental finding or a pure hypothesis) and may identify some of the

possible mechanisms of influence by identifying all biologically significant links B that have been reported in the literature such that A influences B and B influences C.

In the past few years we have designed a tool, ARROWSMITH, which we believe can help investigators to formulate and assess their scientific hypotheses. We have previously discussed its methodological aspects [9] and have published several examples in which ARROWSMITH searches have provided information of interest to neuroscientists [3–8]. Here we present an informal tutorial intended for potential users of the simplified experimental version of ARROWSMITH that is now freely accessible on the World Wide Web (http://kiwi.uchicago.edu). While the details of conducting ARROWSMITH searches are covered in a User's Guide on the Web site, this paper is concerned with discussing the overall rationale and with identifying reasons that a user may wish to run such a search in the first place.

2. Materials and methods

We begin at the point where an investigator seeks to assess the possible relationships that may exist between two items, A and C, which may represent the names of physiologic parameters (e.g. blood pressure), diseases, drugs, nutrients, proteins, etc. A relation between A and C may already be established by experimental, epidemiologic or genetic linkage studies, or the user may simply wish to hypothesize that A and C are related in some manner. Two MEDLINE searches are carried out by the user, creating an A file consisting of all titles containing the term A (and synonyms and alternative spellings, if appropriate), and a C file consisting of all titles containing the term C. Then the ARROWSMITH software simply compiles a list of all words and phrases B common to the two sets of titles. Many of the B-terms will be predictably non-interesting (e.g. 'the,' 'patient') and are excluded using a precompiled stop-list of 5000 words. One can also apply a more stringent criterion, if desired, so that items will appear on the B-list only if they appear in two or more titles in both the A and C sets. The investigator then scrutinizes the B-list, looking for items that he or she feels might plausibly link A and C.

For each such B-term that is identified, another pair of lists are compiled, to obtain an AB list of titles that contain both the terms A and B, and a BC list of titles that contain both B and C. By juxtaposing these lists, the investigator can quickly judge whether the titles indicate a biologically significant relationship linking A, B and C. Such a relationship may, in fact, be well-known in the field, even if it was not known by the investigator doing the search; in such cases, ARROWSMITH may be regarded as providing a type of information retrieval that extends the basic capability of MEDLINE searches. However, those hypotheses which appear to be strongly supported by the literature, yet have not been previously tested experimentally, may represent undiscovered public knowledge.

3. Results and discussion

The wide range of potential applications of ARROWSMITH can be appreciated by giving specific examples.

3.1. Suggesting new therapeutic approaches

An early ARROWSMITH search explored possible ways in which Mg may be linked to migraine [7]. At the time this analysis was performed, no prior papers in the literature had tested whether there was a biologically-significant relationship between Mg and migraine. Yet 11 B-terms existed whose AB and BC titles were each strongly suggestive of such a relationship. Moreover, these linkages were consistent: one would expect a local or systemic deficiency of Mg to worsen migraines, and Mg supplementation to prevent or ameliorate migraines. After Swanson published a paper bringing attention to this hypothesis [7], 12 papers appeared reporting experimental or clinical

tests, and all but one were confirmatory [9]. An analysis of citation patterns of these papers suggests that the authors of these studies were stimulated to carry out these tests, at least in part, from having read the earlier ARROWSMITH paper [8].

Dietary fish oil was linked to Raynaud's disease in a similar fashion [6]. Multiple B-terms indicated that a set of physiologic parameters were well-known to be abnormal in Raynaud's disease, and dietary fish oil was well-known to alter the same parameters in a direction which should normalize the abnormalities in the disease. Yet at the time the ARROWSMITH search was performed, no papers had examined whether dietary fish oil would ameliorate the signs and symptoms of Raynaud's disease. After Swanson published a paper bringing attention to this possibility [6], a clinical trial carried out subsequently by independent investigators has confirmed the beneficial effects of fish oil on patients with the disease, and citation patterns indicate that the clinicians had read the earlier ARROWSMITH paper [8].

3.2. Anticipating adverse drug reactions

For example, anti-inflammatory agents such as indomethacin are currently being examined for their utility in treating or preventing Alzheimer's disease. Though the primary action of indomethacin is clear (inhibiting prostaglandin synthesis), indomethacin has so many effects in so many organs that it is difficult to anticipate which, if any, effects of indomethacin might be expected to be of special relevance to patients with Alzheimer's disease. We carried out an ARROWSMITH analysis with A = indomethacin and C = Alzheimer's disease, looking for terms B that might plausibly link these items [3]. Although most of the B-terms that we found appeared to represent beneficial effects of indomethacin, terms such as 'acetylcholine' drew our attention to the fact that indomethacin has been reported to inhibit effects of, or release of, acetylcholine in a variety of peripheral and central neurons. Since Alzheimer's patients are thought to have decreased cholinergic activity within the cerebral cortex that contributes to their dementia, this may be regarded as a potential adverse drug reaction, and those involved in clinical trials with indomethacin should at least be alert to this possibility [3].

3.3. Identifying mechanisms by which bioactive compounds modulate cellular or organismal responses

For example, clinical and epidemiologic data indicate that estrogen has beneficial effects on human memory, and may decrease the risk of Alzheimer's disease. Yet estrogen affects so many targets, in so many organs, that it is difficult to make a systematic list of the targets that are especially likely to be relevant to Alzheimer's disease. An ARROWSMITH analysis with A = estrogen and C = Alzheimer's disease generated a list of 194 proteins or physiologic parameters modulated by estrogen, including eight plausible candidates which had previously been investigated in the context of Alzheimer's disease, yet had not been explored as possibly explaining estrogen's protective effects on this disease [4]. The eight items represent a 'wish list' of highly plausible targets to explain estrogen's protective effects, that have not appeared in published articles already, but which investigators may be expected to pursue in the near future. Scientists routinely make use of the same information-gathering strategy in deciding which experiments to pursue next, but ARROWSMITH facilitates this task by automatically and systematically creating a comprehensive list of B-terms to be considered.

3.4. Identifying potential animal models for human disorders

A recent paper reported that a novel calcium-independent form of phospholipase A_2 is selectively elevated in the serum of patients with schizophrenia [10]. This finding is provocative, yet it is difficult to assess its significance. One would like to have an animal model for identifying the cellular origin of this enzyme, studying its normal regulation, and learning

which situations may elevate its levels in the serum. Such an animal model does not exist, to our knowledge, but ARROWSMITH can assist in identifying related models that have already been described in the literature. An ARROWSMITH analysis of A = calcium-independent phospholipase A_2 and C = schizophrenia revealed an intriguing B-term, vitamin E, which drew our attention to the fact that a model of chronic oxidative stress in rats (produced by combined vitamin E-selenium deficiency) has been reported to cause a selective elevation of calcium-independent phospholipase A_2 in several tissues [11,12]. This suggests that it is worthwhile testing whether serum levels of this enzyme are also elevated selectively in this animal model. If so, and if the enzyme represents the same one that has been described in schizophrenics, it may represent a promising model for understanding how a specific feature of schizophrenia may arise and be regulated [5].

In conclusion, ARROWSMITH aids biomedical investigators by extending the basic search capability of text-based databases such as MEDLINE, in a manner that facilitates hypothesis formation and assessment. ARROWSMITH does not employ 'artificial intelligence,' but rather filters and juxtaposes information to make it easy for humans to exert their own expert judgment. By implementing a demonstration version of ARROWSMITH in freely available form at http://kiwi.uchicago.edu, we hope to stimulate its further development as a routine informatics tool for the scientific community.

References

[1] D.R. Swanson, Undiscovered public knowledge, Library Q. 56 (1986) 103–118.
[2] M.C. Capone, D.M. Gorman, A. Zlotnik, Identification through bioinformatics of cDNAs encoding human thymic shared Ag-l/stem cell Bg-2. A new member of the human Ly-6 family, J. Immunol. 157 (1996) 969–973.
[3] N.R. Smalheiser, D. R. Swanson, Indomethacin and Alzheimer' s disease, Neurology 46 (1996) 583.
[4] N.R. Smalheiser, D.R. Swanson, Linking estrogen to Alzheimer's disease: an informatics approach, Neurology 47 (1996) 809–810.
[5] N.R. Smalheiser, D.R. Swanson, Calcium-independent phospholipase A_2 and schizophrenia, Arch. Gen. Psychiatry (1998) in press.
[6] D.R. Swanson, Fish oil, Raynaud's syndrome, and undiscovered public knowledge, Perspect. Biol. Med. 30 (1986) 7–18.
[7] D.R. Swanson, Migraine and magnesium: eleven neglected connections, Perspect. Biol. Med. 31 (1988) 526–557.
[8] D.R. Swanson, Intervening in the life cycles of scientific knowledge, Library Trends 41 (1993) 606–631.
[9] D.R. Swanson, N.R. Smalheiser, An interactive system for finding complementary literatures: a stimulus to scientific discovery, Artif. Intell. 91 (1997) 183–203.
[10] B.M. Ross, C. Hudson, J. Erlich, J.J. Warsh, S.J. Kish, Increased phospholipid breakdown in schizophrenia: evidence for the involvement of a calcium-independent phospholipase A_2, Arch. Gen. Psychiatry 54 (1997) 487–494.
[11] C.-F. Kuo, S. Cheng, J.R. Burgess, Deficiency of vitamin E and selenium enhances calcium-independent phospholipase A_2 activity in rat lung and liver, J. Nutr. 125 (1995) 1419–1429.
[12] J.R. Burgess, C.-F. Kuo, Increased calcium-independent phospholipase A_2 activity in vitamin E and selenium-deficient rat lung, liver and spleen cytosol is time-dependent and reversible, J. Nutr. Biochem. 7 (1996) 366–374.