The Arrowsmith Project: 2005 Status Report

Neil R. Smalheiser

UIC Psychiatric Institute, University of Illinois-Chicago, MC912,
1601 W. Taylor Street, Chicago, IL 60612 USA
neils@uic.edu

Abstract. In the 1980s, Don Swanson proposed the concept of "undiscovered public knowledge," and published several examples in which two disparate literatures (i.e., sets of articles having no papers in common, no authors in common, and few cross-citations) nevertheless held complementary pieces of knowledge that, when brought together, made compelling and testable predictions about potential therapies for human disorders. In the 1990s, Don and I published more predictions together and created a computer-assisted search strategy ("Arrowsmith"). At first, the so-called one-node search was emphasized, in which one begins with a single literature (e.g., that dealing with a disease) and searches for a second unknown literature having complementary knowledge (e.g. that dealing with potential therapies). However, we soon realized that the two-node search is better aligned to the information practices of most biomedical investigators: in this case, the user chooses two literatures and then seeks to identify meaningful links between them. Could typical biomedical investigators learn to carry out Arrowsmith analyses? Would they find routine occasions for using such a sophisticated tool? Would they uncover significant links that affect their experiments? Four years ago, we initiated a project to answer these questions, working with several neuroscience field testers. Initially we expected that investigators would spend several days learning how to carry out searches, and would spend several days analyzing each search. Instead, we completely re-designed the user interface, the back-end databases, and the methods of processing linking terms, so that investigators could use Arrowsmith without any tutorial at all, and requiring only minutes to carry out a search. The Arrowsmith Project now hosts a suite of free, public tools. It has launched new research spanning medical informatics, genomics and social informatics, and has, indeed, assisted investigators in formulating new experiments, with direct impact on basic science and neurological diseases.

1 Introduction

In the 1980s, Don Swanson proposed the concept of "undiscovered public knowledge." He published several examples [1-4] in which two disparate literatures (i.e., sets of articles having no papers in common, no authors in common, and few cross-citations) nevertheless held complementary pieces of knowledge that, when brought together, made compelling and testable predictions about potential therapies for human disorders. I was conducting neuroscience research and teaching a course on "The Process of Scientific Discovery" at University of Chicago in the early 1990s

A. Hoffmann, H. Motoda, and T. Scheffer (Eds.): DS 2005, LNAI 3735, pp. 26 – 43, 2005.

when I got a phone call from Don, asking me if I could explain an apparent artifact in his recent analysis of "magnesium" as a term that was pervasive in the neuroscience literature. Everywhere he looked, no matter what neurological disease, it seemed that magnesium was implicated! I explained that this was no artifact – indeed, glutamate excitotoxicity and flow of calcium ions through the NMDA glutamate receptor were thought to be fundamentally important in neurological pathophysiology in a wide variety of conditions, and magnesium was an endogenous factor that controlled the permeability of this receptor to calcium ions. This brief phone call led to a collaboration that now has stretched well over a decade. In the '90s, we published a number of literature-based predictions together [5-8] and created a computer-assisted search strategy ("Arrowsmith") for literature-based discovery [9, 10]. Don then created a free, public demonstration website for conducting Arrowsmith searches, though perhaps the turning point in evolution of the project occurred when Ron Kostoff of the Office of Naval Research asked us to conduct a one-year pilot study to test whether Arrowsmith searches could be used to assist intelligence officers in gathering and integrating disparate pieces of information [11, 12].

These experiences suggested that Arrowsmith might be ready for testing among a wider audience of investigators. Since I am a neuroscientist, it was natural to focus on the biomedical community, but at the same time, it was not clear whether most bench scientists wanted or needed the kind of information that Arrowsmith could provide. Would they find routine occasions for using such a sophisticated tool? Furthermore, as of 2001, it took many hours to carry out a single search, including crafting Arrowsmith queries, navigating the website, and analyzing the results. Would typical biomedical investigators be sufficiently motivated to learn to carry out Arrowsmith analyses? Would they uncover significant findings that affect their experiments or suggest new research directions? Even more uncertain was the question of which, if any, funding agency would support development and testing of Arrowsmith. Fortunately, Stephen Koslow of NIMH had spearheaded a unique NIH-wide program called the Human Brain Project, which sought to establish an informatics infrastructure for neuroscientists to pursue a new paradigm of scientific investigation – one which does not simply formulate self-contained hypotheses, but integrates concepts across disciplines and across investigators. His philosophy was that it is not enough to acquire new data; rather, scientists must be able to carry out data mining, data sharing and data re-use [i.e. re-analyze previous experiments done by others]. The Human Brain Project issued grants that were also unique – each funded program had to combine neuroscience research and informatics research – and they showed enthusiastic support for the Arrowsmith project.

Our early publications had emphasized the so-called one-node search, in which one begins with a single literature (e.g., that dealing with a disease) and searches for a second unknown literature having complementary knowledge (e.g. that dealing with potential therapies). However, we soon realized that the two-node search is better aligned to the information practices of most biomedical investigators: In this case, the user chooses two literatures A and C and then seeks to identify terms B that occur in the titles of both literatures, that point to meaningful links between them. Thus, when in 2001 we initiated a so-called Phase I grant to demonstrate feasibility of the Arrowsmith search tool, we focused almost exclusively on the two-node search strategy. Marc Weeber, an enthusiastic, visionary informatics researcher [13, 14],

gave us crucial assistance at the start of the project; and Vetle Torvik, a brilliant and creative young mathematician [15, 16], has joined us as Project Manager.

2 Project Aims

The Specific Aims of the project are as follows:

1. To test whether Arrowsmith analyses are feasible and useful for assessing research issues, in field tests of neuroscientists working as part of large multi-disciplinary groups; and to incorporate feedback from these users to improve the implementation of the Arrowsmith software.
2. To test whether incorporating MEDLINE record fields other than titles in Arrowsmith analyses will enhance its ability to analyze biomedical literatures.
3. To test whether the free Arrowsmith web site, once upgraded and redesigned with new instructional material, can be made a feasible and useful public forum for conducting Arrowsmith analyses.
4. To test whether Arrowsmith analyses can facilitate inter-laboratory and cross disciplinary collaboration, by identifying complementary sets of investigators that may benefit from working together.

It is most appropriate to discuss our progress on Aims 1-3 together, since our experience with field testers was a major factor in developing both the Arrowsmith website at UIC and the underlying back-end databases. Don has continued to maintain and improve the original website (http://kiwi.uchicago.edu), which has both two-node and one-node search capabilities. However, this site requires users to learn how to upload and download files from the biomedical literature in a particular format, and is limited to 10,000 articles in each set. To overcome these limitations, we created a separate non-mirror site for two-node searches (http://arrowsmith.psych.uic.edu) which is fully interoperable with the popular PubMed interface operated by the National Center for Biotechnology Information (NCBI) – users simply carry out two separate literature searches using the familiar PubMed interface, then click a button and receive a B-list on the webpage within a minute or two. To accomplish this, a dedicated server was set up to handle multi-user queries at UIC; a local copy of MEDLINE was imported, parsed and regularly updated; and the underlying Arrowsmith software was written with fully documented, optimized Perl code. We also programmed a simple, intuitive user interface that field testers found easy to navigate without the need for a tutorial.

The current implementation of the two-node search requires the user to conduct two separate PubMed queries A and C, which define two corresponding sets of articles A and C. (Often the best strategy is to search for articles in which the query term appears in the title. However, the system makes no restriction on how the PubMed queries are conducted; they may involve use of Medical Subject Headings or affiliation fields, may be restricted to review articles, etc.) The Arrowsmith software then stems the titles of the papers in each literature, and makes a list of all single words and two- and three-word phrases that are found in common in the titles of both literatures. Terms that are on a stoplist (consisting of the most common and nonspecific words) are removed, and terms that appear only once in a literature are

removed automatically if the literature is larger than 1000 records in size. The resulting "raw B-list" is then filtered and ranked further before being displayed to the user.

Although the field testers obtained a raw B-list quickly and easily, they found it daunting to analyze because for large literatures, there may be hundreds to thousands of terms on the list even after stoplisting. Therefore, a major research effort has been devoted to developing filtering and ranking schemes, so that users can be shown a single short ranked list of the ~50 most promising terms that match their query needs. As summarized below, we have utilized 7 different filtering methods to create a single ranked list. We have two search tracks on the website: a Basic Search option, in which nearly all filters have preset default settings, and an Advanced option in which the user can set filter settings at will.

Filter 1 involves pre-mapping all of the terms (words and up to 3 word phrases) from MEDLINE titles through the National Library of Medicine MetaMap program [17] to identify those that map to one or more semantic categories as defined by the Unified Medical Language System, UMLS [18]. Then, users can examine only those B-terms that fit into one or more desired categories. Because MetaMap cannot optimally recognize terms out of context and because the UMLS is incomplete for some terms (especially protein and gene names), we have added the Tanabe-Wilbur list of predicted gene and protein names [19] as a back-up to identify this semantic category more accurately.

Filter 2 is a frequency filter; users examine B-terms that occur more than (or less than) a certain number of times in either the A or C literature.

Filter 3 is a recency filter; users can examine B-terms that appeared for the first time more recently than (or only earlier than) a given year in either literature.

Filter 4 incorporates information from MEDLINE Medical Subject Headings (MeSH). For each B-term, the MeSH headings of the AB and BC papers are examined [excluding the 20 most frequent MeSH terms in MEDLINE from consideration]. If they have no MeSH terms in common (or fewer than a threshold number), the B-term is removed from the list (Swanson et al., MS submitted for publication).

Filter 5 merges highly related terms within the same semantic category into a single composite B-term, using a statistical model of term co-occurrence within papers (title or abstract fields).

Filter 6 employs "characteristic terms" calculated for the A and C literatures, which are terms in title or abstract fields that occur in that literature significantly more often than in MEDLINE as a whole. Terms that are not characteristic in either literature are removed as being unlikely to be especially significant.

Filter 7 involves cohesion of B-terms: We hypothesize that for any two B-terms, all other things being equal, the one that represents the more narrowly focused literature will be more useful to the user. Thus, we have defined a measure of term cohesion (based on the set of articles in MEDLINE that contain the term in title; Swanson et al. MS submitted) and have pre-calculated cohesion scores for all terms found in MEDLINE titles. When displayed to the user, the most cohesive B-terms are ranked highest and the lowest may be discarded from the list entirely.

3 Field Tester Experiences

Field testers contributed to the development of Arrowsmith in a number of ways. They documented their own "spontaneously conceptualized" two-node searches in entries made in an electronic notebook that contained the details of the search including the query, all edits made and the final B-list, which were sent to UIC. The entries gave the underlying rationale for the search, and rated the ability of the two-node search to find useful information. We also followed-up to track how often the information affected the course of their research in terms of new ideas (that enriched the discussion section of papers or grant proposals), new experiments conducted and new discoveries made. Two 2-day orientation/training sessions were given at UIC (slides of the lectures can be viewed at the Arrowsmith site), and we also visited the field test sites to give demonstrations and lectures. It is important to mention that the field testing sites consist of active neuroscience investigators whose ongoing work generates diverse types of data, ranging from electrophysiology to brain imaging to microarrays to microscopic tissue sections. Moreover, each site is also engaged in their own neuroinformatics research projects. Thus, field testers are in an ideal position to suggest ways that Arrowsmith searching can be adapted specifically in different ways to meet the diverse needs of neuroscientists and other investigators.

We found that even experienced, oriented, well-trained biomedical investigators did not pursue either simple PubMed searches or Arrowsmith searches in the manner envisioned and advocated by information scientists. Most users typically were looking for one or a few, recent papers on a specific topic, and looked for these on the first page of retrieved titles [i.e. the top 20 hits ranked chronologically by PubMed]. They had no interest in finding ALL relevant papers comprehensively. Nor did they attempt to craft their queries carefully -- rather, the queries were deliberately or casually underdetermined and they expected to sift through some irrelevant papers as they scanned quickly through the first few retrieved pages. Their definition of success is quite different from that of an information scientist, and neither recall nor precision (not even top-20 precision) are attuned to this strategy. We are still not quite sure whether this "Googling" strategy is one of naivete or shrewdness!

Field testers did, indeed, find that Arrowsmith assisted them in assessing hypotheses and identifying promising experiments to pursue, and we are currently writing up a paper that will document these findings in detail. A number of the hypotheses assessed by Arrowsmith searches have grown into new research collaborations. However, we found that field testers employed Arrowsmith routinely for three other tasks as well: a) Many of the searches posed for Arrowsmith could have been pursued with simple PubMed searches, but were found convenient to conduct within the Arrowsmith web interface – for example, some users entered two separate searches A and C when desiring to search "A AND C." b) In addition to seeking conceptual links between two literatures, users often wanted simply to construct a list of items studied in both literatures. c) Many searches were conducted by a user familiar with field A who wanted to browse within an unfamiliar literature C, hoping to find a subset of articles in C relevant to A. We plan to tailor the Arrowsmith interface in the future to support these needs (see below).

4 Progress Towards Facilitating Collaboration

In contrast to Aims 1-3, Aim 4 was considered exploratory, and over the last four years a variety of eclectic approaches have been pursued to develop tools and venues that can assist investigators in finding and facilitating potential collaborations.

To disambiguate author names in MEDLINE, we created a quantitative model, **"Author-ity,"** to estimate the probability that any two papers (sharing the same author last name and first initial) are authored by the same individual [20]. This approach, which will be described below in further detail, has grown into a project to cluster all papers in MEDLINE according to author-individuals. Having such information should assist investigators in finding potential collaborators, particularly in conjunction with planned efforts to create author profiles and to map collaborative networks among people publishing in MEDLINE.

Explicit guidelines were also formulated to help two academic investigators negotiate collaborations [21].

We recognized the need for a forum to create connections among biomedical investigators, who need information management and data integration tools; CS/ informatics researchers, who devise these tools; and social scientists, who evaluate these tools in the context of scientific practice. To create such a forum, a new peer reviewed, open access journal is being launched at BioMed Central, called **Journal of Biomedical Discovery and Collaboration**. I serve as Editor-in-Chief, and William Hersh of Oregon Health Science University is Deputy Editor.

Finally, we invited a prominent social informatics researcher, Carole Palmer of Univ. Illinois-Champaign-Urbana, to analyze the broader information needs and practices of neuroscientists in the field testing sites in detail (participation by field testers was voluntary, but most participated enthusiastically). This project was funded separately as a 3-year project by NSF; the analyses were designed, carried out and are being written up by the UIUC team [22] independently of the Arrowsmith Project.

5 Extending Arrowsmith in New Ways and New Databases

5.1 Incorporating Abstract Terms as B-terms in MEDLINE Searches

Perhaps the single most often-asked question we hear regarding Arrowsmith is, "Why are B-terms taken only from titles?" One reason is that titles of MEDLINE articles are usually very informative, and this maintains a much higher signal-to-noise ratio than simply using terms taken willy-nilly from abstracts and full-text. Another reason is that we already obtain hundreds to thousands of title B-terms from a typical two-node search based on titles. However, there are a number of reasons to include terms in the abstract and full-text of papers as well. First, the title conveys only a small portion of the total information contained in a scientific paper [e.g., 23-25]. Second, terms appearing in the title of a paper may play a qualitatively different role than terms appearing in the abstract – for example, when the term "calpain" appears in the title of a paper, the paper is likely to be studying calpain itself (its enzyme activity, its gene expression, its substrates, etc.). In contrast, when papers contain "calpain" only in abstracts, the authors may be using calpain or calpain inhibitors as experimental

reagents. Third, we have found that users often want to examine a B-list, not only to find meaningful conceptual links between two different literatures, but also to quickly construct a list of items that are in common to the two literatures – such items might potentially include affiliations, funding sources, methods, etc., that never appear in the titles of the papers. Finally, there are other approaches to analyzing two literatures that do not involve constructing a B-list for the user at all (see browsing display, below), which may make use of terms in abstract and full-text.

Therefore, a major thrust of our current research is to examine how best to incorporate terms from abstract and full-text into the two-node search. The challenge will be to devise appropriate strategies of restriction, filtering and ranking that elevate the most useful B-terms above a very significant potential "noise" level. These are not simply programming tasks, but involve issues both of basic informatics – how to represent and mine the textual information – and of aligning the display and logical flow of the search process to the needs of scientific users. Handling abstract terms is, in many ways, a straightforward extension of the approaches already outlined for title terms. In contrast, the issue of extracting terms from full-text involves mining PubMed Central rather than MEDLINE, is associated with different user queries, has a number of special problems related to heterogeneity of text portions within an article, and will require establishing a modified user interface.

The basic method of identifying B-terms within abstracts is to scan the abstracts for all terms that are expressed in both A and C literatures. However, if this is performed with no further restrictions, the size of the raw B-list will be an order of magnitude larger than that obtained using title terms alone, and the number of AB and BC papers per B-term will also increase several-fold on average. Therefore, we will only consider terms that occur in at least 2 article titles within MEDLINE (terms that do not occur in titles are mostly broken phrases that arise from stemming artifacts). Then, the list of "raw" B-terms will be filtered and ranked as for B-terms obtained from titles. Even so, we expect that many incidentally mentioned terms will survive the current filtering process, so one or more additional restrictions will be tested and implemented: **1**. B-terms may only be chosen if they occur in the final sentence of the abstract, or within the Conclusion section of a structured abstract. The last sentence often summarizes the main finding of the paper, so this should give the maximal signal-to-noise ratio within the abstract. **2**. In addition to choosing terms from the final sentence, B-terms may also be chosen if they co-occur in the same sentence as a term that occurs in the title of the same paper. **3**. B-terms may also be chosen only if they are characteristic terms of either A or C literatures. Terms that are not characteristic in either A or C literatures are unlikely to convey important, specific information across the disciplines. **4**. When one is interested in identifying terms that may indicate previously unreported links, B-terms that are characteristic in both literatures are probably already well studied. Thus, as a user-specified option for certain purposes, B-terms may only be chosen if they are "characteristic terms" in one literature but <u>not</u> both.

The Arrowsmith tool is generic, and to date we have deliberately refrained from constraining the type of terms or the type of search that a user can perform. Because of this, two persons carrying out the same search on A = "calpain" vs. C = "postsynaptic density" could have entirely different goals in mind: One could be looking for a list of calpain substrates that are located in the postsynaptic density,

whereas another could be looking for a list of proteases (other than calpain) that cut postsynaptic proteins. However, an important class of search involves looking for statements that "A affects B" or "B interacts with C." Several tools, including MedMiner [26], Chilibot [27] and BioIE [28] employ part-of-speech information and utilize information regarding interaction verbs to identify sentences that discuss interactions between entities, or even specific types of entities such as genes or proteins. Similar NLP techniques may be employed to assist in identifying relevant B-terms within abstracts.

5.2 Incorporating the Use of Full-Text Terms as B-terms

A local copy of PubMed Central (PMC) can be obtained from NCBI, which includes all papers that are publicly accessible (a.k.a. "open access"). PMC full text articles are XML formatted in a standard manner, and can be parsed to create a database capturing each distinct tagged section in the paper (title, abstract, authors, affiliation, introduction, methods, figure legends, tables and table legends, results, discussion, conclusion, acknowledgments, references). Heuristics do need to be developed to recognize sections in cases where sections are not explicitly tagged (e.g., some journals do not label the Introduction as such, some articles lump Results and Discussion together, and so on). Within each tagged section, text can be split into sentences, so that each sentence will comprise a distinct database entry to facilitate searching of term pairs that co-occur within sentences. Each paper can also be cross-referenced to the paper's PubMed descriptors and (if the paper is also indexed in MEDLINE) to the information encoded in MEDLINE fields. PubMed Central contains 372,000 items as of May 2005, with a total of ~800,000 items expected by the end of 2005.

At least initially, users will employ a separate search interface for full-text queries in PMC. Users will specify two different queries that define two literatures A and C, and the user will be asked to specify a particular type of information to be obtained from a menu of choices:

A) Certain types of information can be processed and presented as B-lists from formally structured fields within the papers, without the need for elaborate filtering and ranking procedures. For example, **author names** common to both literatures can be readily identified from the author field; **affiliations** shared in both literatures, terms used in **acknowledgments** (which may include funding sources and thank-you's to colleagues not named as authors), and d) references cited in both literatures (i.e., **co-citations**).

B) Certain types of information are presented in a more variable form but can be recognized by simple look-up: For example, **reagents or assays** described in methods sections, **anatomical regions** or **diagnostic procedures** mentioned in figure legends, or **genes** listed within tables. Users can specify both the section(s) of the paper to be examined, and the semantic category or nature of the B-term desired from a menu. We can identify the vast majority of such terms by simple look-up, using our existing lists of terms that are mapped to UMLS semantic categories.

C) Finally, we plan to tackle the problem of identifying terms within full-text that can supplement the use of title and abstract terms for making conceptual links across the two literatures. We will only consider terms that occur in at least 2 titles in

MEDLINE (see above), and will avoid sections of the paper such as introduction, methods and discussion where many incidental or historical mentions may occur. Thus, B-terms will only be taken from titles, abstracts, figure legends, tables and table legends, and results (though the user will have the option to add or subtract sections from the list). Additional restrictions will also be implemented, similar to those discussed above for abstract terms in MEDLINE, prior to applying filtering and ranking procedures. Depending on how many instances of a single B-term per paper may be found, and how inclusion of full-text terms affects the size of the B-list, it may be necessary to cluster, compress and/or summarize the sentences related to the same B-term for display to the user.

5.3 Current Challenges

There are at least five limitations in the filtering and ranking of B-terms. **First**, filtering B-terms by UMLS semantic categories does not have ideal flexibility – e.g., one can specify the category of receptor, but not restrict it further to NMDA receptors. **Second**, ranking B-terms by coherence values gives undue weight to very rare B-terms, and we are still learning how to correct this properly. **Third**, the problem of word-sense disambiguation has not been addressed yet. **Fourth**, the default stoplist is approximately 8200 words, which was originally chosen manually by Don Swanson with some further editing at UIC. Because it is difficult to be sure that all of these words are predictably non-interesting to all potential users, we plan to construct a smaller and more rationally chosen stoplist, and we are exploring whether words having extremely low coherence can be fruitfully added to that list. Alternative 1400 and 365 word lists of the most common words are already available within the advanced search settings. **Fifth**, there are both advantages and disadvantages to tokenizing the terms prior to processing (stemming, stoplisting, removing uppercase and splitting into sentences). This speeds processing greatly in two-node searches, but does not allow NLP analyses of free text. At present, the local copy of MEDLINE contains titles represented both in original and tokenized versions; however, the abstracts are only saved in tokenized form, and our term database consists only of tokenized terms. Therefore, new databases will need to be created if we decide to employ information gathered from analyses of free text such as part-of-speech tagging or parsing.

5.4 Adding an Alternative Display for Cross-Disciplinary Browsing

A surprisingly common, yet previously unanticipated, reason that field testers employed the Arrowsmith two-node search was to browse in an unfamiliar discipline, looking broadly for articles that might be relevant to one's home discipline. In this situation, scrutinizing a B-list is more of a distraction than an aid in identifying relevant papers. Assuming that the user is familiar with literature A, and that the user is not familiar with non-overlapping literature C, the goal is to identify a subset of papers within C that is most closely related to A. Previous studies have employed ontologies or customized standard vocabularies to connect literatures [29]. Certainly MeSH terms could also be used for this purpose. However, MeSH terms may not be ideal for connecting literatures that deal with basic science rather than clinical

medicine, and particularly may be too limited in the case of very disparate literatures. For example, shared MeSH terms might not be useful for linking "pesticides" with "fast Fourier transforms." Eventually one would like to be able to connect biomedical articles to literatures found in other fields entirely, such as agriculture, psychology, education and engineering. Thus, just as we have chosen to employ shared B-terms to connect disparate literatures for regular two-node searches, so do we hypothesize that the articles most relevant for scrutiny by a browsing user will be BC, the subset of C that shares certain B-terms with A. Three sub-problems need to be solved in order to create the browsing mode:

1. The size of the subset BC must be chosen so that it represents a relatively small proportion of the C literature. Using the entire raw B-list to define BC would result in a set almost as big as C itself. However, using the default filtering and ranking scheme of the regular two-node search (semantic category, frequency, MeSH and cohesiveness filters) is a promising approach: Using the top 50-100 ranked B-terms results in BC subsets that typically contain only ~200-500 papers. The size of the BC subset is closely linked to the number of B-terms but is relatively insensitive to the size of C. Thus, choosing the top 100 B-terms for a large literature (e.g. schizophrenia, which contains over 50,000 papers) results in a BC subset that represents about 1% of the total. Choosing the optimal size and filtering of the B-list for defining BC is an empirical problem (rather than a theoretical one) and will require trial-and-error testing over a variety of specific searches.

2. A method for clustering the BC papers by topic, and giving a short label to each cluster, needs to be chosen and implemented. Although numerous methods have been explored for thematic clustering, the requirements of clustering in the present context create a number of specific constraints: a) The method needs to be computationally quick, so that it can be computed for thousands of articles within a few seconds. b) Clusters should ideally be "soft"; that is, if individual papers fit several clusters equally well, they can be placed in both. c) The clusters should fit well with the user's conception of how the literature is coarsely organized according to topics. d) Last, but not least, the clusters should be viewable by the user on one webpage. Once the user chooses one cluster, it can be displayed and then optionally re-clustered into another set of subclusters, thus permitting drilling-down of the literature in hierarchical fashion.

The "Anne O'Tate" utility (a separate feature of the Arrowsmith website designed to allow simple data-mining of literatures) currently makes use of MeSH headings in a set-covering approach to form clusters within a set of articles retrieved from PubMed in the following manner: First, all MeSH headings mentioned in the article collection are listed in descending order of frequency. The MeSH that occur in >1/3 of papers are deemed less useful for grouping subclusters, so they are bypassed; for the most frequent MeSH term (below 33%), all papers indexed by that term [and any MeSH terms below that term in the MeSH hierarchy] are placed in cluster #1 and removed from the stack. The MeSH term frequencies are re-calculated for the remaining papers, and the process is repeated to form cluster #2, and so on, until either the clusters contain only single papers or 15 clusters have formed. Any remaining papers [and any papers not indexed with MeSH headings at all] are placed in a final cluster called "other." Finally, for each cluster, a new query is performed containing the original query AND the specific MeSH term defining that cluster – this

retrieves the additional papers indexed by that MeSH which had been placed earlier into other clusters, so that individual papers are placed into multiple clusters where appropriate. This method is fast, robust, soft and intuitive, and immediately gives an annotation for the cluster (namely, the MeSH term used to define it).

5.5 Revisiting the One-Node Search

In the Arrowsmith one-node search [9], we begin with a single starting literature A (e.g., the literature on migraine), and compile a list of all terms B_i in the titles of this literature. The list of B_i terms is filtered using a stoplist, and in some cases is further filtered to keep only terms that occur in literature A significantly more than in MEDLINE as a whole. For each B_i term, we search MEDLINE for all papers having B_i in the title [as a practical restriction on the search space, these searches are often restricted to articles sharing a certain MeSH term, such as "Pharmacologic Actions"]. The set of all articles found by searching term B_i is called literature C_i, and all terms in the titles of these papers are referred to as C_i terms. The C_i terms are filtered to keep only terms that occur in literature B_i significantly more than in MEDLINE as a whole, and in some cases, C_i terms are removed if they occur in literature A at all. Then, the C_i terms are combined across all B_i searches to form a master list of C terms. These are ranked according to the number of distinct B terms with which they co-occur -- the presumption is that high ranking C terms are likely to point to previously undocumented, yet biologically meaningful relationships with the A literature.

Many information scientists have explored refinements to the original one node search strategy: Gordon and colleagues used lexical statistics [30] and Latent Semantic Indexing [31] to identify other literatures that contain complementary information. Weeber used UMLS concepts (rather than text words) captured in full-text of articles [13]. Hristovski and Srinivasan have used MeSH terms rather than text words [32, 33], and Wren used manually-constructed "objects" and used a mutual information measure [34-36] to rank objects according to the strength of linkage across literatures A and C. Others, including Pratt [37] and Hearst [38], have explored ways to enhance the user interface to support one-node searching. The Arrowsmith Project offers a public one-node search interface at http://kiwi.uchicago.edu, and Hristovski (http://www.mf.uni-lj.si/bitola/) and Pratt http://litlinker.ischool.washington.edu/) maintain search websites as well, which indicates a high level of interest in these services.

Nonetheless, we deliberately did not study one-node searches as part of the field tester experiences. One concern was that the typical biomedical scientist might not give sufficient credibility to the findings of a one-node search – the indirect links found in the structure of the biomedical literature do not necessarily correspond to the structure of nature itself! Another concern is that the one-node search is an exercise in "searching for an hypothesis," whereas most scientists already have more hypotheses than they can handle, and instead want a tool [the two-node search] to help them assess the ones they already have. Finally, although one-node searches have led to significant testable biomedical predictions, none of the proposed means of filtering and ranking C terms have undergone theoretical or empirical validation. Yet the one-node search can be viewed in several respects as a variant or refinement of the

two-node search, and we believe that the time is ripe for tackling the one-node search again.

For example, consider the typical two-node search, in which the user specifies two topical PubMed queries that define literature A and C. One could, instead, input a topical query for literature A [say, "microRNA"], and allow literature C to be all of MEDLINE, or to correspond to a broad MeSH category such as "Disease". This would create an asymmetric situation similar to that envisioned in a one-node search. The current Arrowsmith interface can support literature sizes up to 100,000 in each query, but as an advanced option the user will be allowed to input files of any size, including all of MEDLINE (actually, handling MEDLINE as literature C is an easy task since we have pre-computed frequencies and PubMed IDs for all terms occurring in MEDLINE titles). Conversely, given any two-node search, we could readily construct a ranked list of C terms that show strong, indirect links with literature A. This will provide a different way of browsing an unfamiliar C literature for items that may be relevant to A, which should complement the article-based browsing approach discussed in the previous section. Thus, both the asymmetrical nature of the search, and the construction of a C-list, can be viewed individually as simple extensions of the existing two-node search.

The issue of how to filter and rank C-terms optimally is not easy, in part because different types of searches may have different optimal strategies. For example, in the case where A_i and C_i refer to specific gene names found in the A and C literatures, it is probably true that if A_i and B_i co-occur often in MEDLINE, and B_i and C_I also co-occur more often than expected by chance, then one would expect A_I and C_i to co-occur as well – and if they do not, this raises the question whether this represents an undocumented discovery of a relationship between A_i and C_i. Thus, for predicting gene interactions, co-occurrence frequencies are probably valid for deriving links. However, in other cases, it is probably not valid to focus only on B-terms that occur more often than expected by chance in both literatures: suppose literature A represents the field of microRNAs, and one seeks complementary information in a disparate field (e.g., nutrition). Rather, the terms most likely to point to undocumented discoveries may be those that are characteristic in one, but not both, literatures.

5.6 Gene-Centric Tools

The basic concept of the two-node search can be extended to other datasets such as those in the GEO gene expression database (maintained, like PubMed and PMC, by NCBI as part of their Entrez suite of databases). Suppose that an investigator hypothesizes that two genes A and C should be co-expressed, but they have not been studied together in previous experiments -- one of the genes may have been discovered recently or is an expressed sequence tag (EST), or the two genes may have been studied in different contexts or species and/or were not included on the same microarrays. A two-node search would allow the investigator to find all genes B that were co-regulated with A in certain experiments and that were, separately, co-regulated with C in other experiments. This would allow one to assess whether a relationship between A and C is likely and warrants further study. Alternatively, suppose an investigator has just made a new lab finding that two genes A and C do

indeed co-express in one situation. It would be valuable to examine the set of other genes (B genes) that have been reported to co-express with both A and C in different experiments. This B-list will assist in placing the A-C relationship into the larger context of gene networks. One may also wish simply to combine data across multiple experiments of the same type. For example, at present, 19 different experiments in the GEO database have examined the expression of eIF2c2 in human brain. (These experiments are not necessarily all comparable to each other, but one could filter them manually if desired.) Making a list of genes that co-express with eIF2c2 across multiple experiments is one way of detecting the genes that are most robustly linked with eIF2c2. Conversely, one might want to compare two apparently disparate experiments and find gene pairs that are co-regulated similarly in both cases.

The GEO Gene Expression database can be adapted for conducting two-node searches in a manner that is analogous to the text-based Arrowsmith search: The user search-interface would replicate the NCBI site (GEO Profiles). A two-node search might go something like this: **1**. The user inputs the name of a gene, together with additional restrictions such as platform, species, tissue or developmental stage. This request is processed to retrieve all GEO Profiles that satisfy the query, giving **literature A**. (A GEO Profile describes a single experiment involving that gene.) Automatically, for each GEO Profile having more than 2 experimental conditions, the software computes all of the "profile neighbors" of that gene – this computation uses the Pearson correlation coefficient to identify a set of the other genes whose expression was most similar to the index gene in that same experiment. Profile neighbors are calculated using a Pearson linear correlation with a threshold of 0.7, and a t-test with an arbitrarily-determined Bonferroni adjustment - the top 100 profile neighbors are presented. **2**. The user inputs the name of a second gene, together with restrictions as desired, which gives **literature C**. The GEO Profiles are retrieved for the second gene, and the "profile neighbors" are computed. **3**. All "profile neighbors" which are common to both literatures are identified and displayed on a single **B-list** of gene names. **4**. The user can select any B-term and see the Profiles that include A and B, juxtaposed to the Profiles that include B and C.

This scheme is quite analogous to the situation of searching two literatures in PubMed. At present, the GEO database is small and extremely heterogeneous, so that it is not easy to formulate useful A and C searches. This limitation should become less important with time, as GEO becomes more populated and as the scientific community formulates standard platforms and standard formats for documenting experiments.

6 Spin-Offs Supported by the Arrowsmith Project

6.1 "Anne O'Tate"

We have programmed a utility on the main Arrowsmith homepage that displays, for any collection of PubMed articles, a ranked list of most frequent terms, most frequent MeSH headings, and most "important" words appearing in title or abstract. The utility also displays most frequent author names, affiliations, journals, a histogram of years of publication, and a list of the terms that have appeared for the first time most

recently in MEDLINE. Also, the user can cluster the articles into topical subgroups (as discussed above). Each of these lists gives a different, partially complementary summary view of the contents of the article collection.

6.2 "WETLAB"

A simple open source electronic laboratory notebook has been programmed in Java, that is oriented to the needs of wet-lab neuroscientists. This notebook, **WETLAB**, allows flexible searching of both data and metadata across templated text fields, stores the text in XML files, and allows data-sharing by ftp or email. WETLAB is currently undergoing beta testing and will be placed on the Arrowsmith website for unrestricted download.

6. 3 Genomics Studies

Unlike the other Arrowsmith projects, this work has not been directed (yet) towards generating a software tool or web service. Rather, we have utilized some of our data-mining approaches to analyze the newly discovered class of genes known as microRNAs and their targets in the mammalian genome [39-43]. This combined computational and wet-lab project involves several of the Arrowsmith field testers and is an important scientific test bed for tool development, as well as an exciting scientific arena in its own right.

6.4 "Author-ity"

As a first step towards creating an author-individual database of all articles in MEDLINE, we have created a statistical model of how two papers authored by the same vs. different individuals vary on a similarity profile computed across different MEDLINE attributes (title words, MeSH, co-author names, affiliation words, etc) [20]. We have programmed a tool, "Author-ity," that resides on the main Arrowsmith homepage: the user specifies a (last name, first name, optional middle initial and optional suffix), and retrieves a list of papers bearing that name. The user then chooses one paper from the list and obtains a ranked list of all of the other papers in descending order of probability that the paper was written by the same individual.

A monotone model is satisfied when the value of the function increases monotonically as the value for a given variable increases (and all other variables' values remain the same). Such functions are easily computed and can place multi-dimensional data onto a single dimensional ranking score or probability value in a manner that takes into account nonlinear and interactive effects across dimensions, yet is readily interpretable for the nature and contribution of each dimension. This type of model appeared to be ideal for the task of comparing two different articles in MEDLINE bearing the same author name, and asking whether they were authored by the same individual.

First, we hypothesized that different papers written by the same individual will tend to share certain characteristic features, not only dealing with the author's personal information (name and affiliation attributes) but other attributes of the articles as well. The probabilistic model [20] describes, for any two papers bearing the same author (last name, first initial), how similar the two papers are across 8

different dimensions: middle initial match, suffix match (e.g., Jr. or III), journal name, language of article match, number of co-author names in common, number of title words in common after preprocessing and removing *title-stopwords*, number of affiliation words in common after preprocessing and removing *affiliation-stopwords*, and number of MeSH words in common after preprocessing and removing *mesh-stopwords*. These are calculated solely from comparing corresponding MEDLINE fields. The resulting 8-dimensional comparison vector, which we call the "similarity profile," is computed for the members of two large reference sets – a match set, consisting of many (millions) pairs of papers very likely to be co-authored by the same individual across MEDLINE, and a non-match set consisting of many pairs of papers known to be authored by different individuals. These training sets were very robust against inclusion of incorrect data.

Thus, given any pair of papers bearing the same author (last name, first initial), we compute the similarity profile and observe its relative frequency in the match set vs. the non-match set. If the observed profile is much more frequent in the match set than in the non-match set, it is likely that the two papers were written by the same individual. The ratio of the profile frequency in the match vs. non-match sets, together with an estimate of the *a priori* probability that any two randomly chosen papers having that name will be authored by the same individual, gives an estimate of the probability that the two papers were written by the same individual [20]. We plan to employ clustering algorithms on papers bearing the same (last name, first initial) to form clusters of papers that can be assigned to distinct author-individuals across MEDLINE.

If monotone models are so wonderful, why aren't they utilized more often in a variety of other situations in medical informatics and bioinformatics, for example, to improve algorithms for information retrieval? Possibly the reason is that it is often hard to generate enough training data to properly fit a monotone model, especially when the number of distinct observable cases is high (e.g., when there are many variables or variables are continuous). Hopefully the use of massive, automatically generated training sets should enhance the popularity of this approach.

7 Conclusions

Don, Vetle and I differ markedly in our backgrounds and personalities, yet are compatible in terms of our general approach to informatics, and this has given a distinct flavor to our joint research efforts:

First, are interested in having computers do what they do best, rather than what people do best. We are not against AI, NLP or machine learning approaches. However, our own goal is to create tools that extend (but not replace) the normal capabilities of people. We seek to make telescopes, not artificial retinas.

Second, we have undertaken a commitment to developing free, public tools. The Arrowsmith websites require no passwords or registration, and although they are under continual development, they are not simply demonstration sites but offer full-strength capabilities for the real-life information needs of scientists.

Third, the tools that we develop are very simple and generic. They are applicable to all fields of biomedical science, by scientists at all levels of seniority, and equally by people running small laboratories or practitioners of Big Science.

Fourth, the field testers are not simply beta testers, experts or "users" but are true scientific collaborators in the development process. It is common in bioinformatics to combine computational biology and wet-lab studies, but I think that the Arrowsmith project has a uniquely multi-disciplinary discovery process that encourages investigators to contemplate radically new directions in their research.

Fifth, we are attuned to a paradoxical requirement of informatics tools: they need to be designed to align well with the perceived needs of scientists and their daily practice, yet the tools also need to be designed to expand scientists' horizons – to improve their ability to handle information and scientific ideas, and to raise expectations and consciousness in a manner that will reshape routine scientific practice [44].

The Arrowsmith Project has demonstrated that it is feasible for scientific investigators to conduct two-node searches in their daily lives. The next challenge is to publicize the tool widely and to induce young scientists, especially, to think explicitly about how they formulate and assess new hypotheses.

Acknowledgments

Supported by NIH grants LM07292 and LM08364. This Human Brain Project Neuroinformatics research is funded jointly by the National Library of Medicine and the National Institute of Mental Health. I thank Marc Weeber, Alan Lian, Wei Zhou, Wei Zhang and Clement Yu for contributions in computer science and informatics, and Amanda Bischoff-Grethe, Lauren Burhans, Christopher Dant, Mike Gabriel, Ramin Homayouni, Alireza Kashef, Maryann Martone, Lauren Penniman, Guy Perkins, Diana Price, Allan Reiss and Andrew Talk for their participation in this collaborative venture as field testers.

Reference

1. Swanson DR. Fish oil, Raynaud's Syndrome, and undiscovered public knowledge. Perspect. Biol. Med. 1986; 30: 7-18.
2. Swanson DR. Undiscovered public knowledge. Library Q 1986; 56: 103-118.
3. Swanson DR. Two medical literatures that are logically but not bibliographically connected. JASIS 1987; 38: 228-233.
4. Swanson DR. Migraine and magnesium: eleven neglected connections. Perspect. Biol. Med. 1988; 31: 526-557.
5. Smalheiser NR, Swanson DR. Assessing a gap in the biomedical literature: magnesium deficiency & neurologic disease. Neurosci. Res. Commun. 1994; 15: 1-9.
6. Smalheiser NR, Swanson DR. Linking estrogen to Alzheimer's Disease: an informatics approach. Neurology 1996; 47: 809-810.
7. Smalheiser NR, Swanson DR. Indomethacin and Alzheimer s Disease. Neurology 1996; 46: 583.

8. Smalheiser NR, Swanson DR. Calcium-independent phospholipase A2 and schizophrenia. Arch. Gen. Psychiat. 1998; 55: 752-753.
9. Swanson DR, Smalheiser NR. An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artif. Intelligence 1997; 91: 183-203.
10. Smalheiser NR, Swanson DR. Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. Computer Methods and Programs in Biomedicine 1998; 57: 149-153.
11. Smalheiser NR. Predicting emerging technologies with the aid of text-based data mining: a micro approach. Technovation 2001; 21: 689-693.
12. Swanson DR, Smalheiser NR, Bookstein A. Information discovery from complementary literatures: categorizing viruses as potential weapons. JASIST 2001; 52: 797-812.
13. Weeber M, Vos R, Baayen RH. Using concepts in literature-based discovery: Simulating Swanson's raynaud - fish oil and migraine - magnesium discoveries. JASIST 2001; 52: 548-557.
14. Weeber M, Vos R, Klein H, De Jong-Van Den Berg LT, Aronson AR, Molema G. Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide..JAMIA 2003; 10: 252-9.
15. Torvik VI, Triantaphyllou E. Guided Inference of Nested Monotone Boolean Functions. Information Sciences 2003; 151: 171-200.
16. Torvik VI, Triantaphyllou E. Discovering rules that govern monotone phenomena. In Data Mining and Knowledge Discovery Approaches Based on Rule Induction Techniques (Triantaphyllou and Felici, eds.) Massive Computing Series, Springer, 2005, Chapter 4: 149-192, in press.
17. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001;:17-21.
18. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med. 1993 Aug;32(4):281-91. Related Articles, Links
19. Tanabe L, Wilbur WJ. Generation of a large gene/protein lexicon by morphological pattern analysis. J Bioinform Comput Biol. 2004 Jan;1(4):611-26.
20. Torvik VI, Weeber M, Swanson DR, Smalheiser NR. A probabilistic similarity metric for MEDLINE records: a model for author name disambiguation. JASIST 2005; 56(2): 140-158.
21. Smalheiser NR, Perkins GA, Jones S. Guidelines for negotiating scientific collaborations. PLoS Biology 2005; 3(6): e217.
22. Palmer CL, Cragin MH, Hogan TP. Information at the Intersections of Discovery: Case Studies in Neuroscience. Proc. ASIST annual meeting. 2004 Nov; 448-455.
23. Kostoff RN, Block JA, Stump JA, Pfeil KM. Information content in MEDLINE record fields. Int J Med Inform. 2004 Jun 30;73(6):515-27.
24. Ding J, Berleant D, Nettleton D, Wurtele E. Mining MEDLINE: abstracts, sentences, or phrases? Pac Symp Biocomput. 2002;:326-37.
25. Shah PK, Perez-Iratxeta C, Bork P, Andrade MA. Information extraction from full text scientific articles: where are the keywords? BMC Bioinformatics. 2003;4:20.
26. Tanabe L, Scherf U, Smith LH, Lee JK, Hunter L, Weinstein JN. MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. Biotechniques. 1999 Dec;27(6):1210-4, 1216-7.
27. Chen H, Sharp BM. Content-rich biological network constructed by mining PubMed abstracts. BMC Bioinformatics. 2004 Oct 8;5(1):147.
28. Divoli A, Attwood TK. BioIE: extracting informative sentences from the biomedical literature. Bioinformatics. 2005 May 1;21(9):2138-9.

29. Chen H, Martinez J, Ng TD, Schatz BR. A concept space approach to addressing the vocabulary problem in scientific information retrieval: An experiment on the worm community system. 1997 Jan. JASIST 48 (1):17-31.
30. Lindsay RK, Gordon MD. Literature-based discovery by lexical statistics. JASIS 1999; 50: 574-587.
31. Gordon MD, Dumais S. Using latent semantic indexing for literature based discovery. JASIS 1998; 49: 674-685.
32. Hristovski D, Peterlin B, Mitchell JA, Humphrey SM. Using literature-based discovery to identify disease candidate genes. Int J Med Inform. 2005 74:289-98.
33. Srinivasan P Text Mining: Generating Hypotheses from MEDLINE JASIST 2004; 55(5): 396-413.
34. Wren JD, Bekeredjian R, Stewart JA, Shohet RV, Garner HR. Knowledge discovery by automated identification and ranking of implicit relationships. Bioinformatics. 2004 Feb 12;20(3):389-98.
35. Wren JD, Garner HR. Shared relationship analysis: ranking set cohesion and commonalities within a literature-derived relationship network. Bioinformatics. 2004 20:191-8.
36. Wren JD. Extending the mutual information measure to rank inferred literature relationships. BMC Bioinformatics. 2004 Oct 7;5(1):145.
37. Pratt W, Yetisgen-Yildiz M. LitLinker: Capturing Connections across the Biomedical Literature. Proceedings of the International Conference on Knowledge Capture (K-Cap'03). p. 105-112. Florida, October 2003.
38. Hearst MA. Untangling text data mining. Proc. Assoc. Comp. Ling. 1999.
39. Smalheiser NR. EST analyses predict the existence of a population of chimeric microRNA precursor-mRNA transcripts expressed in normal human and mouse tissues. Genome Biology 2003; 4:403.
40. Smalheiser NR, Torvik VI. A population-based statistical approach identifies parameters characteristic of human microRNA-mRNA interactions. BMC Bioinformatics 2004;5:139.
41. Smalheiser NR, Torvik VI. Mammalian microRNAs derived from genomic repeats. Trends in Genetics 2005; 21(6): 322-326.
42. Smalheiser NR, Torvik VI. Complications in mammalian microRNA target prediction. To be published in "MicroRNA: Protocols", ed. S.-Y. Ying, in the series "Methods in Molecular Biology", Humana Press, 2005.
43. Lugli G, Larson J, Martone ME, Jones Y, Smalheiser NR. Dicer and eIF2c are enriched at postsynaptic densities in adult mouse brain and are modified by neuronal activity in a calpain-dependent manner. J. Neurochem. 2005, in press.
44. Smalheiser NR. Informatics and hypothesis-driven research. EMBO Reports 2002; 3: 702.