DESCRIPTION: State the application's broad, long-term objectives and specific aims, making reference to the health relatedness of the project. Describe concisely the research design and methods for achieving these goals. Avoid summaries of past accomplishments and the use of the first person. This description is meant to serve as a succinct and accurate description of the proposed work when separated from the application. If the application is funded, this description, as is, will become public information. Therefore, do not include proprietary/confidential information. **DO NOT EXCEED THE SPACE PROVIDED.**

Conventional search engines (e.g., PubMed) can identify many facets of information contained in Medline records and give an indication of the current state of knowledge on a given topic. However, such queries are insufficient for investigators operating at the frontiers of scientific discovery, who need to assess the possible significance of new findings or hypothesized relationships that have not yet been experimentally tested. "Arrowsmith" is a computer-assisted strategy designed to address this problem by facilitating the identification of information that is present implicitly within or across databases. It does this by identifying title words that are shared across a pair of disjoint literatures, filtering out words that are likely to be non-informative, and justaposing titles of papers in each literature that share a given title word, allowing the user to assess whether the two sets of papers are suggestive of a biologically meaningful inference when considered together. To test the feasibility of Arrowsmith for use by the scientific community, a set of field testers will be trained to evaluate the feasibility and utility of Arrowsmith analyses when used in the context of multi-disciplinary neuroscience groups that have both significant laboratory and informatics components. The utility of linking papers via shared terms taken from other fields of Medline records, e.g., abstract words or MeSH subject headings, will also be evaluated. Access to Arrowsmith will be provided via a Web site that will be upgraded and expanded to accommodate the larger public as well. Finally, it will be assessed whether Arrowsmith analyses can help facilitate new interdisciplinary research by identifying and alerting complementary groups of investigators who independently are tackling problems of joint interest, that may be best addressed by working together. Because of the generality of the Arrowsmith approach, it may be applied to a wide range of biomedical problems and, indeed, to conduct searches within or across any databases that contain textual material.

PERFORMANCE SITE(S) (*organization, city, state*)

Primary site: University of Illinois at Chicago, Department of Psychiatry, College of Medicine, 1601 W. Taylor Street M/C 912, Chicago, IL 60612.
Subcontract sites:
University of Illinois at Champaign/Urbana, Beckman Institute, 405 N. Mathews Avenue, Urbana, IL 61801.
University of Chicago, Division of the Humanities, 1010 E. 59th Street, Chicago, IL 60637.
University of California at San Diego, Department of Neurosciences, San Diego, CA 92093.
Stanford University, Department of Psychiatry, 401 Quarry Road, Stanford, CA 94305.

KEY PERSONNEL. See instructions on Page 11. *Use continuation pages as needed* to provide the required information in the format shown below.

Name	Organization	Role on Project
Neil R. Smalheiser, MD, PhD	University of Illinois at Chicago	P.I.
TBA	University of Illinois at Chicago	Project Manager
TBA	University of Illinois at Chicago	Info. Scientist/Programmer
Michael Gabriel, PhD	University of Illinois at Champaign/Urbana	Subcontract P.I.
TBA	University of Illinois at Champaign/Urbana	Postdoctoral Field Tester
Don R. Swanson, PhD	University of Chicago	Subcontract P.I.
Maryann Martone, PhD	University of California at San Diego	Subcontract P.I.
TBA	University of California at San Diego	Postdoctoral Field Tester
Allan Reiss, MD	Stanford University	Subcontract P.I.
TBA	Stanford University	Postdoctoral Field Tester
Stephan Eliez, MD	Stanford University	Co-Investigator

Type the name of the principal investigator/program director at the top of each printed page and each continuation page. (For type specifications, see instructions on page 6.)

RESEARCH GRANT

TABLE OF CONTENTS

*Type density and type size of the entire application must conform to limits provided in instructions on page 6.

Appendix (*Five collated sets. No page numbering necessary for Appendix.*) [X] Check if Appendix is included

Number of publications and manuscripts accepted or submitted for publication (*not to exceed 10*) 5

Other items (*list*):

BIOGRAPHICAL SKETCH

Provide the following information for the key personnel in the order listed on Form Page 2.
Photocopy this page or follow this format for each person.

NAME	POSITION TITLE
Neil R. Smalheiser	Assistant Professor, Dept. of Psychiatry

EDUCATION/TRAINING (*Begin with baccalaureate or other initial professional education, such as nursing, and include postdoctoral training.*)

INSTITUTION AND LOCATION	DEGREE (*if applicable*)	YEAR(s)	FIELD OF STUDY
University of Iowa, Iowa City, IA	BA (Honors)	1974	Mathematics/Gen. Sci.
Albert Einstein College of Medicine, NY, NY	MD-PhD	1982	Medicine/Neuroscience

RESEARCH AND PROFESSIONAL EXPERIENCE: Concluding with present position, list, in chronological order, previous employment, experience, and honors. Include present membership on any Federal Government public advisory committee. List, in chronological order, the titles, all authors, and complete references to all publications during the past three years and to representative earlier publications pertinent to this application. If the list of publications in the last three years exceeds two pages, select the most pertinent publications**. DO NOT EXCEED THREE PAGES.**

1982-1983	Wyler Children's Hospital, Chicago	Internship
1983 -1985	Univ. of Chicago Dept. of Pediatrics	Postdoctoral work
7/1/86-6/30/88	Univ. of Chicago Dept. of Pediatrics	Instructor
7/1/88 - 7/1/96	Univ. of Chicago Dept. of Pediatrics Committee on Developmental Biology Committee on Neurobiology, and the College	Assistant Professor
9/1/96-present	Univ. IL at Chicago Dept. of Psychiatry Dept. of Anatomy/Cell Biology	Assistant Professor Adjunct Assistant Professor

HONORS AND AWARDS

Ford Future Scientists of America Regional Award, 1968. National Merit Scholarship Program Finalist, 1971. B.P.O. Elks Scholarship, 1971. Honors Scholarships, University of Iowa, 1971-1973. Phi Beta Kappa, 1972. B.A. with Honors (General Science) and graduation with High Distinction, 1974. NIH Medical Scientist Training Program Fellowship, 1974-1981. Committee on Pathobiology Training Grant, Univ. of Chicago, 1983. NIH NRSA individual postdoctoral training award, 1984-1985. Andrew W. Mellon Foundation Fellow, 1988-1989. Who's Who in Science and Engineering, 1991.

Block Fund grant, 1986. Brain Research Foundation grants, 1984-87; 1993. Dysautonomia Foundation grants, 1986-88. Schweppe Foundation career development award, 1987-1990. March of Dimes Basil O'Connor Starter Scholar award, 1987-1989. March of Dimes grant, "Laminin as a Molecular & Genetic Probe of Neurites," 1990-1992. NIH NS FIRST Award, "Molecular and Cellular Basis of Cranin's Action on Neural Cells," 1988-1994. NIH HD Program Project, "Biological Basis of Mental Retardation," P.I. of Project #2, 1992-1995. Scottish Rite Schizophrenia Research Program, "Heat shock protein 60 serum antibodies in schizophrenia," 1993-1995. Office of Naval Research, "ARROWSMITH analysis of biomedical innovation and discovery," 1999-present. NIMH R03, "Circulating Reelin and Psychosis Vulnerability," 2000-2002.

PEER REVIEWING ACTIVITIES

Behavioral and Brain Sciences; Brain Research; Journal of Biological Chemistry; Journal of Cell Biology; Journal of Clinical Investigation; Journal of Neurochemistry; Journal of Neuroscience; Journal of Neuroscience Research; Life Sciences; Mech. Aging and Development; Proceedings of the Society of Experimental Biology and Medicine; PNAS; Restorative Neurology and Neuroscience; National Science Foundation (Developmental and Cellular Neuroscience); U.S.-Israel Binational Science Foundation; Basic Science Foundation (Israel Academy of Sciences and Humanities).

PROFESSIONAL ORGANIZATIONS

American Association for the Advancement of Science; Society for Neuroscience; American Society for Cell Biology; International Society for Neurochemistry; International Society for Developmental Neuroscience; Associate, Behavioral and Brain Sciences; Licensed physician in the State of Illinois.

RESEARCH PROJECTS ONGOING OR COMPLETED DURING THE LAST 3 YEARS:

"ARROWSMITH Analysis of Biomedical Innovation and Discovery." PI: Neil R. Smalheiser. Office of Naval Research / Dept. of Human Systems Science and Technology (Medical Division). Period: active, 8/1/99 – 7/31/00. ARROWSMITH is a novel computer-assisted bioinformatics strategy. The purpose of this grant to assess the value of ARROWSMITH analyses in areas relevant to Defense priorities. We have been invited to submit a proposal for continuation of this project for another year, consideration of which is pending.

"Innovation and Discovery from Complementary Literatures." Co-Investigator: Neil R. Smalheiser. Department of Defense / Defense Intelligence Agency. Period: active, 8/1/99-7/31/00. The PI on this grant is Don R. Swanson, PhD, of the University of Chicago. This grant is being carried out as a collaboration with the one above, albeit each is granted to a different institution and each emphasizes different aspects of Arrowsmith analyses.

"Circulating Reelin and Psychosis Vulnerability," PI: Neil R. Smalheiser. NIMH R03. Period: 7/1/00-6/30/02. Proposes to establish a RIA for quantifying circulating reelin, and to raise antibodies against C-terminal domains of reelin as a preliminary step towards studies in humans. The goal is to assess the value of circulating reelin as a marker of psychosis vulnerability.

"Reelin and Brain Function in Neuropsychiatric Disease," PI: Neil R. Smalheiser. NIMH R01. Proposes to analyze the cellular and molecular mechanisms by which reelin is secreted, processed, and bound to the extracellular matrix in developing and mature mammalian brain. Submitted June 1, 1999 – will be resubmitted after additional preliminary studies are completed. Currently, exploratory investigations are underway to assess whether reelin is abnormal in CSF and plasma of schizophrenic patients, to investigate the role of reelin as a trophic factor within extra-CNS tissues, and to elucidate basic biochemical mechanisms of reelin action on target cells.

None of these projects, nor any of the subcontractors' support, have any overlap with the present proposal.

SELECTED PUBLICATIONS (from the last 10 years)

Smalheiser, N.R. (1990) Neuronal growth cones: An extended view. Neurosci. 38: 1-11.

Smalheiser, N.R. (1990) Cell attachment and neurite stability in NG108-15 cells: Effects of 5'-deoxy,5'-methyl thioadenosine (MTA) compared with laminin, kinase inhibitor H-7, and Mn^{2+} ions. Devel. Brain Res. 51:153-160.

Smalheiser, N.R. (1991) Cell attachment and neurite stability in NG 108-15 cells: What is the role of microtubules? Devel. Brain Res. 58: 271-282.

Smalheiser, N.R. (1991) Role of laminin in stimulating rapid-onset neurites in NG 108-15 cells: Relative contribution of attachment and motility responses. Devel. Brain Res. 62: 81 - 89.

Pomeranz, H. D., Sherman, D. L., Smalheiser, N. R., and Gershon, M. D. (1991) Expression of the immunoreactivity of a neurally related cell surface laminin binding protein by neural crest-derived cells migrating to and within the gut: relationship to the formation of enteric ganglia. J. Comp. Neurol. 313: 625-642.

Smalheiser, N.R. and Collins, B.J. (1992) Characterization of a novel set of membrane antigens associated with axonal growth. I: Biochemical and functional studies. Devel. Brain Res. 69: 215-223.

Smalheiser, N.R. and Collins, B.J. (1992) Characterization of a novel set of membrane antigens associated with axonal growth. II: Expression in the chick central nervous system. Devel. Brain Res. 69: 225-231.

Smalheiser, N.R., Collins, B.J. and Sharma, S.C. (1992) Characterization of a novel set of membrane antigens associated with axonal growth. III: Expression in the regenerating goldfish optic nerve and tectum. Devel. Brain Res. 69: 277-282.

Landis, C., Collins, B., Cribbs, L., Sukhatme, V., Bergmann, B., Rechtschaffen, A., and Smalheiser, N. R., (1993) Regional expression of Egr-1 in the brain of sleep deprived rats. Mol. Brain Res. 17: 300-306.

Smalheiser, N.R. (1993) Acute neurite retraction elicited by diverse agents is prevented by genistein, a tyrosine kinase inhibitor. J. Neurochem. 61: 340-343.

Smalheiser, N.R. (1993) Cranin interacts specifically with the sulfatide-binding domain of laminin. J. Neurosci. Res. 36: 528-538.

Smalheiser, N. R. and Ali, J. Y. (1994) Acute neurite retraction triggered by lysophosphatidic acid: timing of the inhibitory effects of genistein. Brain Res. 660: 309-318.

Smalheiser, N. R. and Swanson, D. R. (1994) Assessing a gap in the biomedical literature: magnesium deficiency and neurologic disease. Neurosci. Res. Commun. 15: 1-9.

Smalheiser, N. R. and Kim, E. (1995) Purification of cranin, a laminin binding membrane protein. Identity to dystroglycan and reassessment of its carbohydrate moieties. J. Biol. Chem. 270: 15425-15433.

Smalheiser, N. R. and Swanson, D. R. (1996) Indomethacin and Alzheimer's Disease. Neurology 46: 583.

Smalheiser, N. R. and Swanson, D. R. (1996) Linking estrogen to Alzheimer's Disease. Neurology 47: 809-810.

Smalheiser, N. R., Dissanayake, S. and Kapil, A. (1996) Rapid regulation of neurite outgrowth and retraction by phospholipase A_2-derived arachidonic acid and its metabolites. Brain Res. 721: 39-48.

Smalheiser, N. R. (1996) Proteins in unexpected locations. Mol. Biol. Cell 7: 1003-1014.

Belkin, A. and Smalheiser, N. R. (1996) Localization of cranin (dystroglycan) at sites of cell-matrix and cell-cell contact: recruitment to focal contacts is dependent upon extracellular ligands. Cell Adhes. Commun. 4: 281-296.

Swanson, D. R. and Smalheiser, N. R. (1997) An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artificial Intelligence 91: 183-203.

Peng, H.B., Ali, A.A., Daggett, D.F., Rauvala, H., Hassell J.R. and Smalheiser, N.R. (1998) The relationship between perlecan and dystroglycan and its implications in the formation of the neuromuscular junction. Cell Adhes. Commun.5: 475-489.

Smalheiser, N. R., Haslam, S. M., Sutton-Smith, M., Morris, H. R. and Dell, A. (1998) Structural analysis of sequences O-linked to mannose reveals a novel Lewis X structure in cranin (dystroglycan) purified from sheep brain. J. Biol. Chem. 273: 23698-23703.

Smalheiser, N. R. (1998) Conserved amphipathic helices near the N-terminus and C-terminus of the alpha subunit of cranin (dystroglycan). Cell Adhes. Commun. 6: 401-404.

Smalheiser, N. R. and Swanson, D. R. (1998) Calcium-independent phospholipase A_2 and schizophrenia. Arch. Gen. Psychiat. 55: 752-753.

Smalheiser, N. R. and Swanson, D. R. (1998) Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. Computer Methods and Programs in Biomedicine 57: 149-153.

Impagnatiello, F., Guidotti, A., Pesold, C., Dwivedi, Y., Caruncho, H., Pisu, M.G., Uzunov, D.P., Smalheiser, N.R., Davis, J.M., Pandey, G.N., Pappas, G.D., Tueting, P., Sharma, R.P., and Costa E. (1998) A decrease in reelin expression as a putative vulnerability factor in schizophrenia. Proc. Natl. Acad. Sci. USA 95:15718-15723.

Swanson, D. R. and Smalheiser, N. R. (1999) Implicit text linkages between Medline records: using Arrowsmith as an aid to scientific discovery. Library Trends 48: 48-59.

Smalheiser, N. R. (2000) Walter Pitts. Perspect. Biol. Med. 43: 217-226..

Smalheiser, N. R., Costa, E., Guidotti, A., Impagnatiello, F., Auta, J., Lacor, P., Kriho, V. and Pappas, G. (2000) Expression of reelin in adult mammalian blood, liver, pituitary pars intermedia and adrenal chromaffin cells. Proc. Natl. Acad. Sci. USA 97: 1281-1286.

Smalheiser, N. R. and Collins, B. J. (2000) Coordinate enrichment of cranin (dystroglycan) subunits in synaptic membranes of sheep brain. Brain Res., submitted.

Majumdar, A., Smalheiser, N. R., Markaryan, A. and Kaplan, A. (2000) Forms of amyloid precursor protein and H-kininogen directly interact in vitro, coimmunoprecipitate from human serum and colocalize in rat brain. J. Biol. Chem, submitted.

BUDGET JUSTIFICATION

The PI (Smalheiser) is responsible for the overall project. The Program Manager will assist in the instruction of field testers, write instructional materials for the Web site, collect and analyze evaluations and feedback from field testers, run and evaluate additional searches on their own, and assist in follow-up efforts to test Arrowsmith suggestions experimentally. One Subcontract PI (Swanson) is responsible for maintaining and upgrading the Arrowsmith Web site, including expanding its capabilities to receive simultaneous multi-user queries, construct B-lists based on shared abstract or MeSH terms, and handle non-Medline databases as warranted. The information scientist/programmer will assist in implementing changes and upgrades to the Web site, and in documenting the Arrowsmith software. The other Subcontract PIs (Gabriel, Martone, Reiss) and the co-Investigator (Eliez) are responsible for recruiting and supervising the postdoctoral field testers, who in turn are responsible for learning and conducting Arrowsmith analyses as outlined in the proposal. Drs. Gabriel and Reiss are not requesting salary because it is fully covered from other resources, whereas the policies of their Universities dictate that Drs. Martone and Eliez request $5,000 for their participation.

Travel is requested for field testers to attend a short course at UIC, and for key personnel to attend the annual meeting of the Human Brain Project (at which time, all project personnel will meet to summarize current progress and problems). As well, funds are requested to allow one visit per year among personnel at UIC and a field test site.

Supplies are requested at each of the sites to support computer searches, and to permit laboratory testing of hypotheses arising from Arrowsmith analyses, either by field testers, or in collaboration with UIC or outside investigators. Because the exact nature of the tests remains to be discovered during the grant period, it is not possible to itemize the supply needs in advance, but are likely to include disposable items such as chemicals, labware and data storage media.

Detailed budgets and justifications are included from each subcontractor (that for Dr. Gabriel is listed below, and the others are on separate sheets following their budgets).

All figures reflect 4% escalation in years 2-5.

Budget justification for Dr. Gabriel:
As with all field test sites, 60% effort (plus fringe benefits) is requested for a postdoctoral field tester to carry out, evaluate and follow up Arrowsmith analyses as outlined in the proposal. Travel is requested for the field tester to attend a short course at UIC, and for key personnel to attend the annual Human Brain Project meeting. Supplies are requested to cover the costs of testing Arrowsmith suggestions. Because the exact nature of the tests remains to be discovered during the grant period, it is not possible to itemize the supply needs in advance, but are likely to include disposable items such as chemicals, labware and data storage media.

RESOURCES

FACILITIES: Specify the facilities to be used for the conduct of the proposed research. Indicate the performance sites and describe capacities, pertinent capabilities, relative proximity, and extent of availability to the project. Under "Other," identify support services such as machine shop, electronics shop, and specify the extent to which they will be available to the project. Use continuation pages if necessary.

Laboratory: (The P.I. occupies a biochemistry wet-lab on the 2nd floor of the UIC Psychiatric Institute. The Institute is well equipped for all modern neuroscience techniques including EM, mass spec, molecular biology, etc. These facilities are not directly relevant to the Arrowsmith proposal, except insofar as they indicate our ability to support possible collaborative studies that may arise from field testers' analyses.)

Clinical: (UIC has a Clinical Research Center which, again, is not directly relevant to the proposal except insofar as it may support collaborative studies identified from field testers' analyses.)

Animal: (UIC has a central animal facility, plus a satellite facility located on the 2nd floor of the Psychiatric Institute for housing mice and rats.)

Computer: The Arrowsmith Web site server, currently located at University of Chicago, will be supplemented by a larger machine to be installed at UIC by March 1st; the latter will be used by the field testers. It will consist of a desktop PC-type computer, Linux operating system, Apache webserver, 20GB storage, 220 MB internal memory, 500MHz processor, and internet connection with permanent IP address.

Office: The P.I. occupies an office next to the lab on the 2nd floor of the Psychiatric Institute. The project manager and information scientist/computer programmer will occupy offices on the 4th floor of the same building.

Other: Subcontract sites:
Univ. of Illinois at Champaign/Urbana – PI has extensive experimental and neuro-informatics facilities within the Beckman Institute, Neuronal Pattern Analysis Group, Neuronal Time Series Analysis Workbench, and National Center for Supercomputing.
Univ. of Chicago – PI has a Sun Sparc5 workstation, and an office on the main campus.
Univ. of CA at San Diego – PI has extensive experimental and neuro-informatics facilities within the National Center for Microscopy and Imaging Research, Natl. Partnership for Advanced Computational Infrastructure and Dept. of Neurosciences.
Stanford University – PI heads Div. of Child and Adolescent Psychiatry, including a Neuroimaging core facility and a group developing software for imaging databases.

a. Specific Aims

One of the exemplary informatics tools available to investigators is the bibliographic database Medline, whose design facilitates the retrieval of information from published papers based on terms explicitly present in the records (e.g., title, abstract, or MeSH taxonomic subject headings). Together with public search engines such as PubMed, Medline has been extremely successful in creating a community of scientists who regularly mine the literature to keep up with current findings and to gain enhanced perspective on their own and others' work. However, we have shown that a significant body of information residing in Medline is not currently accessible by conventional online searching: this consists of implicit information which is not stated explicitly in the records associated with any single paper, but rather can be surmised only by examining pairs or groups of papers together (1,2). In preliminary studies, we have developed a computer-assisted strategy of data mining that we call "Arrowsmith," and have demonstrated that Arrowsmith can identify information that is contained implicitly, but not explicitly, within Medline (2). Such information can be useful in assessing new scientific findings and hypotheses, particularly those that bridge different disciplines.

To date, our own group have been the primary users of Arrowsmith, and it remains to be seen whether this strategy can have general applicability among the scientific community. We hypothesize that Arrowsmith can be developed into a major data mining tool for Medline, and possibly other databases as well. To do so, one must show that neuroscientists can be taught to conduct and interpret Arrowsmith analyses after a few days of instruction; that they encounter situations amenable to Arrowsmith analyses on a routine basis; and that the results of Arrowsmith analyses do, indeed, prove useful in their hands in assessing new scientific findings and hypotheses. In order to test these hypotheses, we aim:

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.

b. Background and Significance

The Human Brain Project is an ambitious attempt to create the neuroscience community of the future -- a community in which scientists do not merely carry out new experiments, but make full use of existing data as a valuable source of information that can be archived, pooled across different investigators, re-analyzed at will, and combined with data from other fields (3). In order to achieve this paradigm shift, three major challenges must be faced: First, databases need to be designed which can archive the unique types of data acquired by neuroscientists. Second, new informatics techniques need to be devised for manipulating and analyzing these data. And third, a climate of inter-laboratory and cross-disciplinary cooperation needs to be fostered.

The present proposal represents a collaboration between a MD-PhD neuroscientist and an information scientist. For the past 7 years, we have worked together to develop a novel data mining strategy, called "Arrowsmith," intended to help biomedical researchers uncover information that is implicit within databases,

particularly information that bridges journal and disciplinary boundaries. (To date, our research has focused on the database Medline, though the strategy could be applied to textual material within or across any databases.) The key idea of Arrowsmith is to identify possible relationships between two different items by choosing appropriate literatures (i. e., sets of Medline records dealing with each item) and then juxtaposing titles of selected groups of papers from each literature. At present, these papers are juxtaposed by virtue of sharing common words or phrases in their titles, because the titles of biomedical papers are generally quite informative with regard to the major experimental findings contained therein. By inspecting the juxtaposed titles, investigators can quickly assess whether, considered together, they suggest information that is not present in any individual paper.

For example, if one group of papers report that drug A increases cyclic AMP levels in heart muscle, and another group of papers report that increasing cyclic AMP levels activates transcription of certain genes in the same system, then together they raise the implicit suggestion that drug A may activate transcription of those genes in heart muscle. Whether this actually occurs, of course, is a matter for experimental testing to decide, but a person trying to assess such a hypothesis would be greatly interested in knowing ahead of time that it is supported by the current literature. Note that this implication would not be retrievable with conventional Medline searches, unless someone had already tested the hypothesis and reported the findings in a single paper.

To date, 8 different examples of implicit hypotheses in the field of neuroscience have been published that were found within Medline using Arrowsmith analyses (2, 4-10). Several of these hypotheses have been subsequently tested and confirmed experimentally or clinically (fish oil/Raynaud's (4) and Mg/migraine (5); see citations and discussion in ref. 11). Although several other methods of creating paths to link papers within the literature have been explored (e.g., 12-14, 30-32), Arrowsmith is arguably the best documented data mining technique that has been reported for textual material, and is the only one that has generated a stream of novel, testable predictions. The earliest published Arrowsmith analysis has been replicated by others using several different informatics techniques (12, 13), and numerous largely favorable reviews of this approach have appeared (15-20). A free, public demonstration web site has been established at http://kiwi.uchicago.edu, which allows anyone to conduct simple, small-scale analyses using files uploaded from PubMed, Ovid and certain Dialog searches. Finally, ongoing studies sponsored by the Office of Naval Research and the Dept. of Defense are applying and extending the scope of Arrowsmith analyses to anticipate future trends in the field of biological warfare. Thus, Arrowsmith has already found widespread acceptance in the field of information science, and its results have attracted some attention within neuroscience and science policy as well. The present proposal reflects our conviction that Arrowsmith should become a routine informatics tool for the scientific community. Nevertheless, both the overall Arrowsmith strategy, and its specific implementation, must first address a number of important critiques and concerns.

For example, some skeptics have doubted that there exists a significant body of implicit information waiting to be mined in databases. The real issue is not whether the implicit information exists per se – the biomedical literature certainly contains an enormous number of links of the form "A affects B" and "B affects C" – but whether many of these represent important implications, and moreover without anyone being previously aware of that fact. We feel that there are at least three common scenarios in which Arrowsmith searches are likely to aid scientists:

a) When individual pieces of the puzzle lie in different disciplines, e.g., nutrition and psychiatry, where few individuals read the literatures of both fields.

b) At the time that a new scientific finding is made, and one needs to assess its significance immediately relative to the current literature.

c) When the person doing the search is not an expert. In our previous publications we have emphasized cases of Arrowsmith analyses that represent "undiscovered public knowledge;" yet this concept does not refer to information that experts do not know privately, but rather to information that cannot be retrieved publicly (1). Our previous publications have, indeed, focused on Arrowsmith analyses that appeared to be newsworthy even

to experts, but most database queries are not intended to seek entirely novel findings. When a scientist conducts a conventional Medline search, he or she is looking for specific information that is news to them, and does not care that someone else in the world may already know that information. Similarly, a scientist may routinely obtain useful information from Arrowsmith analyses even if it is already "known" to certain experts, or even discussed in the text of certain papers but not retrievable from Medline records.

Some skeptics have raised the opposite criticism – there are likely to be so many implicit links in the literature of the form "A affects B" and "B affects C", that it may be difficult for a human user to pick out the most interesting implications systematically or comprehensively. At best, conducting Arrowsmith analyses may be an art, depending primarily on intuition and pattern recognition in the mind of the person doing the search. We agree that Arrowsmith does not replace human judgment, but argue that it filters and juxtaposes two literatures in a way that minimizes "noise" and maximizes the ability of humans to discern true patterns. Ultimately, the only way to counter this criticism is to examine empirically the process and outcome of Arrowsmith searches when carried out by a number of different users, as proposed here. Another possible objection is that most scientists today have little training in informatics and barely utilize the power of conventional Medline searches. Thus, they lack competence in constructing pairs of complementary literatures, and may lack the patience needed to wade through long lists of B-terms and juxtaposed titles. We believe that the requisite training can be conveyed by short courses of a few days' duration, and again, assessing this question is part of the present proposal.

There are, indeed, certain technical limitations in the current implementation of Arrowsmith software that need to be addressed to make this data mining technique more efficient and user-friendly. For example, the software does not currently match up title words that are related, though not identical, due to alternate word usage (e.g., cell death vs. cell survival, or smallpox vs. variola, or sAPP vs. protease nexin II). Conversely, some words have more than one meaning (lead as a verb vs. lead as a substance), or are investigated in different contexts in different literatures (e.g., E may represent vitamin E in one literature and apo E in another). Most of these limitations should be overcome by incorporating the information provided by the MeSH indexing terms, so that papers in each literature which share common MeSH terms are juxtaposed (either as an alternative to title word grouping, or as a means of further filtering the output provided by title word grouping); this will be explored in Specific Aim 2.

Nevertheless, there will always remains the problem of deciding whether implications that are found are biologically plausible and worthy of further attention. In the case discussed above, where drug A increases cyclic AMP levels in heart muscle, and another group of papers report that increasing cyclic AMP levels activates transcription of certain genes in the same system, the implication is relatively strong. However, if drug A raises cyclic AMP only slightly or transiently, or if the activation of genes was studied only in some other cell type, species, or developmental stage, then the implication is less certain. Regardless of how Arrowsmith is implemented, there is an inherent trade-off between maximizing the chance of finding new potential biologically relevant relationships, and the risk of suggesting false-positive relationships.

In conclusion, Arrowsmith is a novel, unique approach to data mining that has the potential to enhance the power of Medline searching, that in the future may also be applicable to other databases, or to linking findings across different databases. More than simply a method of information retrieval, we believe that Arrowsmith should aid scientific progress at the frontiers of knowledge, by helping scientists at the stage of hypothesis formation and testing. As the size of Medline increases in the future, Arrowsmith should become even more efficient and useful. Moreover, Arrowsmith analyses hold the promise of actively facilitating collaboration across laboratories and across disciplines, by identifying different investigators that may benefit from working together. For example, we have pointed out that a group of psychiatric investigators studying elevated calcium-independent phospholipase A_2 in the serum of schizophrenics (21) might benefit from examining an animal model of oxidative stress reported by a group of nutritionists, in which calcium-independent phospholipase A_2 is

elevated in multiple tissues (22) (ref. 10; see also Preliminary Studies). Instead of each group continuing to work separately, it would be most efficient if the latter group could simply provide serum to be tested by the former group. As a final component of the present proposal, we plan to devote resources specifically to notifying investigators identified by Arrowsmith analyses and assessing whether this can be effective in encouraging the formation of new collaborations.

c. Preliminary Studies/Progress Report

General expertise. Neil R. Smalheiser, MD, PhD, is a neuroscientist with broad experience in studying molecular, cellular and developmental aspects of neurite behavior and the role of extracellular matrix proteins in neural function and dysfunction. As a member of the UIC Psychiatric Institute, he is part of a multi-disciplinary group of basic and clinical scientists with major interests in the neurobiologic basis of psychotic and affective disorders. Don R. Swanson, PhD, formerly Dean of the Graduate Library School at the University of Chicago, is one of the pioneers of the field of information science. Beginning 15 years ago, he formulated the concept of "undiscovered public knowledge," and in collaboration with Dr. Smalheiser has devised the collection of techniques that are subsumed under the Arrowsmith strategy.

Arrowsmith methodology, and examples of Arrowsmith analyses of particular interest to neuroscientists. The underlying rationale, the mechanics of searches, and the criteria for evaluating the outcome of searches have all been extensively discussed in previous publications emphasizing methodology:

- Swanson, D. R. (1986) Undiscovered public knowledge. Library Quarterly 56: 103-118.
- Swanson, D. R. (1987) Two medical literatures that are logically but not bibliographically connected. J. Am. Soc. Inf. Sci. 38: 228-233.
- Swanson, D. R. (1989) Online search for logically-related noninteractive medical literatures: A systematic trial-and-error strategy. J. Amer. Soc. Inf. Sci. 40: 356-358.
- Swanson, D. R. (1990) Medical literature as a potential source of new knowledge. Bull. Med. Libr. Assoc. 78: 29-37.
- Swanson, D. R. (1990) The absence of co-citation as a clue to undiscovered causal connections. in Borgman, C. L., ed. Scholarly Communication and Bibliometrics (Sage Publ., Newbury Park, CA). 129-137.
- Swanson, D. R. (1991) Complementary structures in disjoint science literatures. in Bookstein, A., Chiaramella, Y., Salton, G. & Raghavan, V. V., eds., SIGIR '91 Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, Chicago, (ACM Press, NY). 280-289.
- Swanson, D. R. (1993) Intervening in the life cycles of scientific knowledge. Library Trends 41: 606-631.
- Swanson, D. R. and Smalheiser, N. R. (1997) An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artificial Intelligence 91: 183-203.
- Smalheiser, N. R. and Swanson, D. R. (1998) Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. Computer Meth. Prog. Biomed. 57: 149-153.
- Swanson, D. R. and Smalheiser, N. R. (1999) Implicit text linkages between Medline records: using Arrowsmith as an aid to scientific discovery. Library Trends 48: 48-59.

Of these, the article in Artificial Intelligence and the tutorial in Computer Methods and Programs in Biomedicine are the most relevant to the present proposal, and are included in the Appendix. Besides methodologic papers, several previous publications have focused on biological implications of specific Arrowsmith analyses (the three most recent examples are included in the Appendix):

- Swanson, D. R. (1986) Fish oil, Raynaud's Syndrome, and undiscovered public knowledge. Perspect. Biol. Med. 30: 7-18.
- Swanson, D.R. (1988) Migraine and magnesium: eleven neglected connections. Perspect. Biol. Med. 31: 526-557.
- Swanson, D. R. (1990) Somatomedin C and arginine; implicit connections between mutually-isolated literatures. Perspect. Biol. Med. 33: 157-186.
- Smalheiser, N. R. and Swanson, D. R. (1994) Assessing a gap in the biomedical literature: magnesium deficiency and neurologic disease. Neurosci. Res. Commun. 15: 1-9.
- Smalheiser, N. R. and Swanson, D. R. (1996) Indomethacin and Alzheimer's disease. Neurology 46: 583.
- Smalheiser, N. R. and Swanson, D. R. (1996) Linking estrogen to Alzheimer's disease: an informatics approach. Neurology 47: 809-810.
- Smalheiser, N. R. and Swanson, D. R. (1998) Calcium-independent phospholipase A_2 and schizophrenia. Arch. Gen. Psychiat. 55: 752-753.

Of these, the most recent example will be summarized, both as a way of demonstrating how a typical Arrowsmith search is conducted, and how it may potentially have an impact on scientific progress. The impetus for the analysis was the appearance of an paper by Ross et al. in Archives of General Psychiatry, which found that levels of phospholipase A_2 are elevated in the serum of schizophrenics, and that surprisingly, the elevation reflects a hitherto unreported calcium-independent form of the enzyme in the serum (21). Such an experimental finding connects two different fields or literatures that were previously thought to be unconnected – namely, calcium-independent phospholipase A_2 and schizophrenia. Can one obtain further insight into the possible mechanism of this effect, or into its possible significance, that could guide further research in this area? Certainly a Medline search combining both terms would come up empty. One could read reviews or confer with experts on calcium-independent phospholipase A_2, but it would be desirable to have a means of focusing quickly on the subset of this literature that is most likely to be relevant to schizophrenia.

Our Arrowsmith analysis began by constructing a literature A consisting of all Medline records that contain the phrase "calcium-independent phospholipase A_2" in the title, and a literature C of all Medline records that contain the word "schizophrenia" in the title. Then, a list B was made of all title words and phrases that are present in both A and C; the list was then automatically filtered by a stop list to remove items that are predictably non-informative (e.g., "the" or "patient"; this list currently contains 7,000 words, and also removes phrases if each of the words in the phrase are on the stop-list), and the list was edited manually to remove additional terms. For each B_i-term that remained, a list of titles containing calcium-independent phospholipase A_2 and B_i was printed and juxtaposed to a corresponding list of titles containing B_i and schizophrenia. From inspecting the juxtaposed lists, we uncovered what appears to be a relevant possible link that bridges the two literatures. Namely, under the B_i-term "vitamin E" were a pair of juxtaposed papers: One, published in the Journal of Nutrition (22), showed that an animal model of chronic oxidative stress (vitamin E and selenium-deficient diet) exhibits elevated calcium-independent phospholipase A_2 in a number of tissues; and the other reported that patients with schizophrenia are deficient in levels of vitamin E relative to controls (23).

Together, these suggested the possibility that chronic oxidative stress, which has been consistently associated with schizophrenia, might underlie the elevation in calcium-independent phospholipase A_2 seen in the serum of patients. This directly suggests two testable predictions: a) calcium-independent phospholipase A_2 should be elevated in the serum of animals subjected to chronic oxidative stress, and b) conversely, treating schizophrenics with antioxidants should reduce their serum levels of calcium-independent phospholipase A_2. We alerted the community via a Letter in the Archives (10), which was published together with a concurring reply by Ross et al (24). It is too soon to learn whether our Letter will have made a difference in the way that the original finding will be followed up; certainly it was successful in informing Dr. Ross for the first time of the existence of this animal model (personal communication).

Validation of methodology. The Arrowsmith methodology has been validated at a variety of levels. At the lowest level, we have taken two overlapping literatures (i.e., that share certain titles in common), subtracted the overlap, and constructed a B-list of title words shared in the non-overlapping literatures. Many of the terms found on the B-list correspond to title words present in the overlap -- or in other words, Arrowsmith successfully "rediscovered" within the nonoverlapping literature many of the relationships that were explicit in the overlap literature (ref. 2 and unpublished observations). More importantly, we have demonstrated numerous examples in which Arrowsmith analyses revealed implicit hypotheses, which were deemed novel enough to be reported in peer-reviewed neuroscience journals (7-10). Furthermore, in at least two early cases (4, 5) there is evidence that others were actively stimulated to test and confirm the hypotheses at least in part because of Swanson's analyses (11). Gordon and Lindsay, using a statistical approach upon complete Medline records to define intermediate literatures, replicated in information-science terms the earliest published example of an Arrowsmith analysis, i.e., that pointed out the possibility that fish oil might ameliorate Raynaud's syndrome (12). Later, Gordon and Dumais (13) used another technique, called latent semantic analysis, to identify intermediate literatures, and again, were able to replicate the finding that fish oil and Raynaud's syndrome were highly, albeit indirectly, connected within the literature. We also demonstrated that an automated version of Arrowsmith could replicate the early trial-and-error finding of an association between magnesium and migraine (2).

An Arrowsmith demonstration web site. A free web site has been established and is being maintained by Don Swanson at http://kiwi.uchicago.edu, which allows anyone to upload literatures A and C from PubMed, Ovid and certain Dialog searches. After a delay of a few minutes, the list of B-terms (filtered through the 7,000-word stoplist) is made available for further manual editing, if desired. After submission of the final B-list, the final output is presented: two lists of titles organized by the common B_i-terms that they share. One can print out each list of titles, or can click back and forth from the AB-list to the BC-list on-screen, allowing one to assess quickly whether implicit information is suggested by the juxtaposed titles when considered together. A separate Web site is being established at UIC using a larger and faster server under the sponsorship of the Department of Psychiatry (see enclosed letter from Dr. Joseph Flaherty, Head, UIC Dept. of Psychiatry), that will be the site utilized by field testers (see Research Design and Methods, below).

B-lists incorporating other fields of Medline records. Ongoing studies sponsored by the Office of Naval Research and the Dept. of Defense employ Arrowsmith analyses as part of a larger team employing informatics to anticipate future trends in the field of biological warfare. One particular task makes of use of methodology that is relevant to the present proposal: to construct a list of viruses which are not currently regarded as threats in biological warfare, yet which – based on findings present in the current literature – suggest that they could become modified or engineered so as to make them credible threats in the future. For this purpose, complementary literatures are chosen according to various biological criteria (e.g., papers describing viruses that have been shown to be stable in aerosols, vs. papers describing viruses whose virulence has been studied using genetic techniques). To identify individual viruses that have been studied in both contexts, it is not adequate to look for shared title words -- since the viruses may not be mentioned in the titles per se. Instead, for each set of papers, the shared MeSH subject headings of the Medline records are listed and examined for headings that correspond to names of viruses. Another issue is to track specific institutes or cities that are engaged in several key areas of research at once; this can be found by creating B-lists based on shared author affiliation fields. Our preliminary results thus strongly suggest that these approaches are fruitful for scientometrics and science policy, and potentially may be helpful in dealing with neuroscience searches as well.

Neuro-Informatics at UIC. Although no organized forum for neuro-informatics has been established at UIC, interest in this subject is very high. For example, Steve Koslow, Mark Ellisman and Bruce Wheeler were all independently invited to give seminars this year, and numerous UIC faculty are conducting innovative research (e.g., Robert Gibbons et al. are developing appropriate statistical methods for analyzing gene expression microarrays, and Dr. Kimi Sugaya has submitted a separate proposal to the Human Brain Project related to

establishing gene expression databases for brain regions). The Psychiatric Institute is sponsoring a "NeuroWeb" site to connect laboratories and programs throughout UIC. Besides teaching medical students and psychiatric residents on neuroscience topics, the PI also teaches graduate students in the Dept. of Anatomy/Cell Biology (College of Medicine) and the Dept. of Biological Sciences (College of Liberal Arts), and has lectured on informatics as part of a Neuroscience Methods core course. Thus, the local environment is quite positive, and it is anticipated that both short courses and traditional semester-long neuro-informatics courses will be developed in the near future. The short courses that will be developed for field testers will also be given at least yearly for interested parties at UIC and elsewhere, and should serve as a catalyst for further efforts in this direction.

d. Research Design and Methods

The goal of the present proposal is to assess the feasibility of Arrowsmith when used under field conditions, and to incorporate feedback into further improvements to the system, so that it can be used widely by the scientific community as a routine informatics tool. However, "feasibility" covers a number of distinct issues that need to be addressed separately over 4 specific aims: Can postdoctoral neuroscience investigators learn to conduct Arrowsmith analyses after only a 3 day short course? Do opportunities for Arrowsmith searches arise routinely in the course of scientific investigation? Are the mechanics of conducting Arrowsmith searches, via the Web site, reliable and user-friendly? Do Arrowsmith searches succeed in identifying implicit information? Will expanding Arrowsmith software capabilities lead to more informative searches, or expand the scope of problems that can be addressed? Does the additional information gained via Arrowsmith help investigators determine their experimental directions? Do suggestions identified in Arrowsmith searches prove to be valid when tested experimentally, and do they lead to significant scientific findings? Do the suggestions help bridge disciplines, and identify complementary sets of investigators that might be brought together to tackle problems of common interest? Each of these steps will advance Arrowsmith incrementally towards the overall goal, and each of these steps will be assessed in the course of field tests separately by persons working largely independently of UIC. Any areas of performance that are sub-optimal should be quickly pinpointed by these methods, so they can be improved though appropriate programming or instruction, and verified via re-testing.

Specific Aim 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.

The overall design of field tests is to recruit postdoctoral neuroscience investigators from large multi-disciplinary groups, educate them in online searching and Arrowsmith analyses, and have them look actively for opportunities to conduct Arrowsmith analyses that arise naturally from the research carried out by their group, or other research which comes to their attention (e.g. at seminars or in the current literature). They will keep a log of their opportunities, and of their actual Arrowsmith analyses: they will conduct the analyses first using "standard" parameters, and then depending on the outcome, will follow up with further analyses using different search parameters. The results of the alternative searches will be rated for user-friendliness, speed, and effectiveness in finding nontrivial implicit information. In turn, the information obtained will be rated for its usefulness as background information, as implying future directions for their own group, or as suggesting future directions that would be more suitable for some other group. The feedback from these analyses, together with specific suggestions for improving the implementation of the software, will be collected and analyzed by the program manager at UIC, and used as input for making improvements as appropriate.

Recruitment. Three field sites have been established as formal subcontractors in this proposal. Each of these represent active multi-disciplinary neuroscience laboratories that already participate in the Human Brain Project or the related NSF initiative, and that deal with a diverse array of experimental techniques and neuro-

informatics databases, including imaging data (Stanford), electrophysiologic data (UIUC) and morphologic data (UCSD). In addition, the Stanley Foundation has agreed to serve as a fourth field test site; this choice should be appropriate since they are actively engaged in analyzing multi-disciplinary data obtained across a large number of investigators. Because the Foundation wishes to maintain their financial and administrative independence, they prefer to be regarded simply as collaborators in this project without drawing a budget (see enclosed letter from Dr. Michael Knable, Medical Director of the Stanley Foundation Research Programs). The PIs of each group will be responsible for recruiting a field tester: this individual is likely to be a postdoctoral neuroscientist who shows general computer literacy and familiarity with Medline (advanced graduate students or assistant professors are also qualified). Despite the costs associated with field testing, we feel that having 4 field test sites are necessary not only to evaluate Arrowsmith seriously, but to set new directions for expanding its scope.

Curriculum of short course. The field testers will convene at UIC for a 3-day course, covering the following topics:

1. A brief overview of Medline. Structure of the Medline database; Medline records; MeSH subject indexing; Venn diagram approach to constructing queries; Boolean logic.
2. Search engines of Medline, particularly PubMed [this is preferred for field testers because it is free, widely available, incorporates MeSH categories into queries routinely, searches dates back to 1966 in a single query, and is linked directly to other Entrez databases]. How PubMed converts user queries into internal queries. Advanced PubMed searches. Use of phrases, truncated terms, synonyms, and field-specific terminology; reiterated searches to refine queries.
3. Hints for efficient searches of Medline; balancing specificity vs. inclusiveness.
4. Other bibliographic databases (EMBASE, PsychInfo, etc.), and other search engines (e.g., Ovid, Medminer) available on the Web or through Dialog. Citation analyses through Web of Science or Scisearch.
5. Practicum -- constructing queries in conventional Medline searches.
6. Basic Arrowsmith concepts: non-interacting literatures; complementary literatures; intermediate literatures; implicit information within databases. How Arrowsmith searches are conducted.
7. Hints for constructing Arrowsmith searches: identify complementary literatures that are not too general nor too specific (~1000-~5000 records each); know the direct overlap of these literatures before going further; subtract overlap before submitting each literature for constructing a B-list.
8. Hints for editing the B-list: the stoplist to filter out automatically many non-helpful B-terms; minimal vs. maximal manual editing to remove additional B-terms that are too general; setting thresholds for B-terms based on the number of papers in which they appear; use of phrases; handling B-terms which are "asymmetric" (i.e., they are found in very few papers in one literature, but many papers in the other literature); setting limits on the number of papers retrieved corresponding to each B-term; subdividing the set of papers retrieved corresponding to each B-term.
9. Hints for juxtaposing papers from each literature that share a common B-term, and for assessing quickly whether these are suggestive of implicit information. Follow up findings with inspection of papers, further Medline searches and/or citation analysis.
10. Alternative methods for juxtaposing literatures: use of abstract words, MeSH subject headings, or author affiliations to form B-lists. Use of keywords derived from semantic indexing or cited papers. Applicability of the Arrowsmith approach to other search engines, other databases and to non-bibliographic information.
11. Practicum and testing – constructing and interpreting several Arrowsmith queries that will be posed to the students. Although it is expected that the requisite facility will be achieved by the end of 2 days, students will be asked to allot 3 days to their visit to UIC in case an extra day proves to be needed.
12. Procedures for keeping logs of Arrowsmith opportunities, for conducting and evaluating Arrowsmith searches, and for tracking use made of the information accrued from Arrowsmith analyses (see below).

Methods of evaluation and incorporating feedback into further improvements - Overview. Each field tester will keep a log of situations encountered that could potentially represent an opportunity to conduct an

Arrowsmith analysis. Typically, these opportunities occur when one makes an experimental or clinical finding that bridges hitherto unconnected disciplines, or when one wishes to assess a novel hypothesis, or when one seeks to retrieve information more deeply than is possible using conventional Medline searches, e.g. when gathering information for writing grant proposals or review articles. It is expected that such opportunities will arise weekly within the purview of their own research, that of others encountered within their multi-disciplinary group, or encountered via reading the current literature. Keeping a log (describing the situation and context of the opportunity) should provide empirical information on the frequency with which Arrowsmith opportunities arise in real life. However, we do not expect every opportunity to be analyzed in detail by the field tester; rather, only those which they deem particularly interesting or promising will be followed up.

It is expected that each field tester will conduct an Arrowsmith analysis on the average of once every 3-4 weeks. It is expected, as well, that each analysis should take about 3 days -- to construct a query, perform the search and analyze the results using "standard" parameters -- plus a few more days for alternative parameters to be examined, and if indicated, for reiteration of the search using new query terms or better definition of complementary literatures. Field testers will document the amount of time spent in each phase. If the analyses prove fruitful in identifying new implicit information or suggesting new hypotheses, additional time will be needed to assess their plausibility and novelty by examining the literature further [inspecting papers identified during the search, doing further Medline searches and citation analyses] and possibly to follow up the results experimentally in the lab. Conversely, those searches that appear to be failures will be re-run and analyzed in particular detail by the project manager, to define exactly what components of the search process need to be addressed. Including time spent in preparing logs, evaluating the searches and communicating feedback to the project manager, it is estimated that each field tester will need to devote roughly 60% effort to the present proposal. It is further expected that this effort will enhance, rather than detract, from other experimental work that the field testers are engaged in -- both by providing them with additional information and with additional pilot funds (see Specific Aim 4) that will help them in testing new hypotheses related to their laboratory studies.

Evaluation of Arrowsmith analyses (logs, search outcomes and suggestions for improvement) across the field testing sites. Initially, each Arrowsmith search will be framed in "standard" parameters, that is: a) The query will be framed in terms of potentially complementary literatures A and C that are between ~1,000 and ~5,000 titles in size, and that have excluded any papers that are present in both literatures. Such sets are neither too small so as to contain few valuable title words (that could be inspected directly without Arrowsmith) nor too big so that the literature is unfocused. (As well, note that the limit for downloading files from PubMed is currently 5,000 records; if larger files need to be handled, Ovid can be used, or several PubMed downloads can be combined.) b) The B-list of terms that are common to both sets, and that are not automatically filtered out by the 7,000 word stoplist, will be severely edited down manually prior to inspection of titles, by removing all terms that appear to be "general" a priori (e.g., amygdala, electrophoresis) in the context of the search query, and retaining only terms that appear to be obviously "substantial" (e.g., these are often names of specific drugs, hormones, receptors, etc.). c) After manual editing of the B-list, for each term B_i, a limit of only the most recent 17 titles in literature A that contain B_i (The "AB_i titles") will be displayed [this occupies roughly a page of text], and juxtaposed with the most recent 17 titles in literature C that contain B_i (the "B_iC titles"). The field tester will then inspect the AB_i and B_iC titles to discern whether the standard search contains any implicit suggestions ("finds"), and will list these, together with his/her evaluation of them (see below). d) Note that the definition of a "standard" search will vary over time, as the display parameters discussed below and the new features discussed in Specific Aim 2 become validated, incorporated into the Web site, and made available to users without requiring any custom specifications.

Such a search should be user-friendly and relatively rapid, but has the risk of missing some important information that may have been present in the B_i-terms that were manually edited out, and in the portion of AB or BC sets that contain more than 17 titles. The field tester will accordingly re-display these specific portions of the output that were initially not displayed, and examine whether any additional information is apparent. This

will allow us to assess whether the extra effort involved in larger searches is likely to be warranted under routine conditions, or whether these parameters should be available only as custom options. Also, we suspect that the most common B_i-terms are NOT more likely to reveal suggestions than those present only once or twice, and want to obtain empirical evidence on this point; therefore, for each "find" made by the field tester we will document how many titles the relevant B_i-terms were associated with in each literature, in order to assess whether those B_i-terms that are present in at least 3 or more titles in each literature are any more likely to reflect biologically-meaningful inferences. Finally, although it is simple to discuss Arrowsmith as juxtaposing literatures A and C via B_i-terms so that inferences of the form "A-B_i-C" can be discerned, in actual practice, we have detected a variety of inferences, including B_i-A-C and others that cannot be easily classified because they involve links made by the investigator among papers that were not directly juxtaposed by the program. To gain empirical evidence on how often "finds" are made that do not fit the classical mold, each "find" will be described as A-B-C, B-A-C or "other".

Each "find" will be evaluated by the field tester and placed in one of 4 categories: Category 1 includes implicit suggestions that appear to be valid, yet obvious or self-evident based on the tester's knowledge of the scientific field being queried, or based on the fact that they are already reported in the overlap literature. Category 2 includes implicit suggestions whose validity and implications for further research cannot immediately be assessed. Category 3 consists of suggestions that might potentially make a difference to their research, and that could be tested in the field tester's group. Category 4 consists of suggestions that appear to be worth following up, but that cannot easily be tested by their own group because they would involve reagents, techniques, expertise or experimental models that are not readily available to them. The search overall will be evaluated by the field tester and placed in one of 4 bins: Bin 1- mechanically unsuccessful search, or too laborious, or failed to make "finds" in Categories 2-4 - overall search not worth the effort; Bin 2- "finds" found in Categories 2-4, but only after expanding the search parameters or reformulating the query; Bin 3- "finds" found in Categories 2-4 readily using standard search parameters; Bin 4- "finds" provided significant new insight or information to the user, or could potentially affect the direction of future research, so that the tester feels that the effort involved in the search was amply rewarded.

Analysis of the evaluation data and expected outcomes. The project manager will collect information from each field tester on a monthly basis: logs, search summaries and evaluations, as well as suggestions for improving the methodology and the Web interface. After about a dozen searches have been recorded, we will begin to accrue statistical information regarding the overall experience of the field testers. If there is a marked disparity among field testers in search outcomes, we will attempt to discern if this reflects differences in subject matter, or differences in strategies of query construction. It is expected that at least 1/3 of searches will make "finds" in Categories 3 or 4. It is further expected that, once field testers are familiar with the mechanics of conducting queries, they will rate few searches as having not been worth the effort. By analyzing these evaluations, as well as explicit feedback from the field testers, we will learn what aspects of the search strategy are most in need of adjustment, whether it be educating the users better in constructing queries, organizing the web site differently, or adjusting search and display parameters. Some of the feedback may indicate the need for expanding the scope of Arrowsmith (e.g., to include additional databases), or pave the way for larger programming projects (e.g., analyzing full-text journal articles) that may lie beyond the scope of the present grant period.

Potential pitfalls. The major pitfall in evaluating Arrowsmith via a small number of field testers is the extensive variability that is likely to be encountered – variability across different searches, across different fields of inquiry, and across different testers. Certain aspects of the system will be easy to assess – e.g., can files be uploaded into the Web site successfully? – but it is much more difficult to evaluate formally whether the information discerned using Arrowsmith searches will directly alter the way that scientists do their research. Perhaps it is well to keep in mind the situation of Medline: the PubMed search engine is extraordinarily successful in terms of its daily traffic, and it usually returns (at least some) nontrivial information even when non-

expert users pose questions to it, yet one would be hard pressed to prove that Medline searches facilitate scientific progress. Another issue is that scientists are conditioned to analyze their problems and plan new experiments in a particular manner, that depends upon reasoning and analogy but that does not traditionally make use of informatics. Thus, initially at least, field testers may feel that Arrowsmith gives them information that they did not ask for, or that does not fit into the framework of their own thinking. We believe this mind-set may be changing; in the field of genomics, for example, investigators can identify and sequence entire new genes strictly on the basis of information that is gleaned from existing database knowledge! However, it remains to be seen whether neuroscientists will view Arrowsmith as a similar adjunct to bootstrapping new findings. Finally, a potential pitfall is that our evaluations do not make use of formal sociological or cognitive psychological assessment tools. However, at present, we feel that our efforts are best directed towards practical issues of feasibility and user ratings of utility.

Specific Aim 2. To test whether incorporating Medline record fields other than titles in Arrowsmith analyses will enhance its ability to analyze biomedical literatures.

This Specific Aim seeks to extend the current capabilities of Arrowsmith software by permitting the creation of B-lists based not on shared title words and phrases, but on different fields of the Medline records -- particularly MeSH subject headings shared words and phrases present in the abstracts of papers. A secondary goal of this Aim is to identify needs for any other types of searches that are posed by the field testers, but that cannot be conducted under the standard query parameters of Aim 1 – for example, searches that require examination of literatures much larger than 5,000 records, or that utilize databases other than Medline.

MeSH. First, we will directly implement improvements to create B-lists based on common MeSH subject headings within Medline records, and make this available as an option on the Web site so that field testers, or the public users of the Web site, can formulate standard queries that take advantage of these expanded features. Besides creating a B-list composed of MeSH subject headings per se, we will also explore the use of MeSH headings to reduce ambiguity among title words and to identify title words that are synonymous. For example, at present, if one creates a B-list of title words, the word "lead" meant as a verb will be linked spuriously to the word "lead" meant as a compound. Conversely, papers that mention "sAPP" will not be linked to those that contain the phrase "protease nexin II," even though these refer to the same protein. Thus, we will explore hybrid criteria that employ both title words and MeSH headings, in order to collect related B_i-terms together and reduce spurious links.

Furthermore, MeSH headings may be helpful in another problem that arises during Arrowsmith analyses. Regardless of the method used to create B-lists, certain B_i-terms are associated with a large number of papers in literatures A or C. These highly used B_i-terms are likely to reflect a currently-appreciated relationship, hence may be suspected to shed little new light during an Arrowsmith query. For example, let us consider the example of a query between estrogen (literature A) and the Alzheimer's disease literature (literature C). The B_i-term "amyloid" occurs in scores of papers within literature C, reflecting an already-proposed pathogenetic relationship, so an investigator conducting a "standard" search would be tempted not to examine such a large, low-yield set of B_iC titles at all. However, there is a risk of missing important novel links that might be apparent from juxtaposing the AB_i papers with a specific subset of the B_iC papers. We believe that a promising way around this problem is to identify the subset of B_iC titles that share MeSH subject headings with the AB_i set (papers having both estrogen and amyloid in the title), thus generating a smaller set of B_iC titles to be examined, that are those most likely to be related in some biological fashion to the AB_i set. Such expanded features will be added to the Web site for those B_i-terms that are associated with a large number of papers (say, 10 or more) in a given literature, and evaluated as part of expanded parameter searches by the field testers.

Abstracts. A question that is frequently raised in discussing Arrowsmith is whether it is worthwhile to create B-lists using words and phrases of the abstracts of papers. A priori, much additional implicit information should be contained in the abstract that is not possible to ascertain by inspecting titles alone. However, we have focused on titles because the titles of scientific papers are generally quite informative, and in particular, when scientists make a strong, consistent finding this often is reflected directly in the title. We suspect that the signal:noise ratio for abstract terms may be unacceptable, since using abstracts will greatly increase the total number of B_i-terms, hence greatly increase the "noise" and the effort needed by the investigator to analyze the resulting output. The same problem would apply to the question of examining terms obtained from full-text searches of papers, when this becomes feasible for a wide cross-section of journals. Moreover, abstract terms are easiest to handle when one is dealing with very small, specific literatures – and yet, in our experience, such queries are better formulated at a more general level that might be conducted using title words alone. For example, one might be interested in relating the neurodevelopmental factor "reelin" to some other literature. Reelin was discovered only 5 years ago, and the literature (including the related term "reeler") consists of only a few hundred titles, which is too small for an effective standard Arrowsmith search. One could create a B-list of words present in the abstract of these papers, but it is probably better simply to phrase the search using title words at a somewhat more inclusive level – e.g. starting with the literature on "neuronal migration" (~1800 records). In sum, we feel that it is preferable to evaluate title terms rather than abstract terms as a basis for Arrowsmith searches by field testers, at least initially.

Nevertheless, it is worth examining the issues involved in handling abstract terms, because looking to the future, it is possible that computer-assisted methods may become available for pre-processing the output so that humans can be given a reasonable set of the most relevant papers to examine. Certainly, one can reduce the number of spurious links that are presented in the output by only choosing abstract (or full-text) terms that are in proximity to a given index term -- either by being present in the same sentence, or being within a specified distance away from each other. Medminer is an example of an existing search engine that fruitfully searches terms or pairs of terms within abstracts that lie in proximity to pre-specified "relationship terms" (25), and use of relationship terms could be employed by Arrowsmith as well. (Although proximity relations cannot be defined using the current PubMed search engine, one can do this via Ovid.)

Therefore, we will implement the creation of B-lists based on shared abstract words and phrases on an experimental basis, and will investigate the consequences of including proximity parameters (i.e., the shared abstract word must be in the same sentence as one of the index words used to define literature A and C, and/or must be within a defined distance of the index words) and relationship terms (i.e., they must be proximal to words that reside on a pre-existing list such as "affects", "causes", etc.) (25), to assess how these parameters affect the size and composition of the B-list relative to the B-list generated by shared title words and phrases. In addition, selected queries carried out by field testers in Aim 1 will be re-run by the program manager to compare the abstract output vs. the title word output, and to assess whether searches might have obtained additional inferences if abstract words and phrases had been considered. Another option that may be explored is to focus on words that not only occur within the abstracts of a given literature, but that are mentioned more than once within the same abstract, since these are likely to be especially meaningful (26). If the results of these explorations look promising, then the capability to construct B-lists using abstract words and phrases will be added to the Web site, and made available to field testers.

Alternative features to be added to the Web site. Swanson is currently extending the capability of Arrowsmith software to construct B-lists using author, author affiliation and journal fields of Medline records with expected support from the Dept. of Defense. Although this work is not considered part of the present proposal, these capabilities will be added to the repertoire of the Web site when the methodology becomes ready. Additionally, some queries may be better mined in databases other than Medline. Because Arrowsmith is a general strategy of data mining, it can technically be implemented to identify implicit suggestions within any database that contains textual information, and indeed can be used to link two different databases. For example,

besides Medline, there are other bibliographic databases (e.g., EMBASE, Scisearch, and PsychInfo; some, but not all databases available through Dialog are already compatible with the Arrowsmith Web site presently), bioinformatics databases with textual annotation (e.g., Entrez Genome, or databases within the Human Brain Project), databases of government technical reports or patents, and even the enormous "database" that is the World Wide Web (28). Alternative methods of generating keywords that can be used as a basis for constructing B-lists include counting token or record frequencies containing shared terms taken across all Medline record fields (12), latent semantic indexing (13), semantic indexing (27) or deriving keywords from the Medline record fields of papers that are cited within a given set of papers (31). The available evidence suggests that these are no more efficient or effective than the basic Arrowsmith approach, and some of the methods are more computationally complex. However, these methods may become relevant to explore in the future, as new opportunities and contexts for Arrowsmith analyses arise. Having a information scientist/programmer should facilitate adapting Arrowsmith to new databases or new B-list strategies to serve the needs of field testers, other groups affiliated with the Human Brain Project, or public users of the Web site. It may not be possible to implement all of the programming that would be desirable within Phase I, but simply identifying areas of need will be extremely valuable in looking towards Phase II studies.

Evaluation. The PI, program manager, field testers, and public Web site users all comprise a set of users who will be posing Arrowsmith queries freely based on real-life neuroscience findings and hypotheses. These queries will serve as a test bed for the new features of Arrowsmith discussed above, and in turn they will provide feedback as to how these new features should best be implemented, and what new features should be developed. For features that will be added to the Web site, the evaluation will take place by field testers as outlined in Specific Aim 1. Swanson will be responsible for studying different ways of constructing B-lists using MeSH headings and abstract words and phrases, and will evaluate the feasibility of these approaches primarily in terms of how it affects the size and composition of the B-list relative to the standard title search. The project manager will contribute to this effort by re-running selected field tester queries to learn how the new features affect the outcomes relative to standard parameters, before adding the features to the Web site and having field testers incorporate the new features directly in their searches. We can refer to a successful track record of informatics-neuroscience collaboration so far, and believe that the same strategy should continue to be successful in moving Arrowsmith forward as new features become added.

Open-ended searches. There is one other important situation in which field testers may formulate queries in Specific Aim 1 that do not appear to be appropriate for "standard" parameters, because they involve uploading files that contain significantly more than 5,000 records. This happens in open-ended searches (called "Procedure I" in ref. 2): for example, when one would like to construct a list of drugs that have NOT been investigated in schizophrenia, but whose properties or effects, investigated in some OTHER context, suggest that they may be worthwhile to examine in schizophrenia. Such a search potentially involves uploading the schizophrenia literature (>40,000 records, literature C) and the literature indexed under pharmaceutical preparations (>200,000 records, literature A), constructing a B-list based on shared title words, and ranking the title words in literature A according to how many different B-words they co-occur with (2). The Arrowsmith Web site does, in fact, have the capability to support open-ended searches at present (referred to on the Web site as Stage 4 and 5). In practice, we have found that most laboratory scientists do not actively seek new hypotheses in an open-ended manner, so we do not plan to evaluate this feature systematically, unless field testers express an interest in this type of search. This mode of Arrowsmith search should not be forgotten, however, since it exemplifies a manner in which informatics can be used as a primary tool to re-analyze experimental findings that may potentially suggest new directions of research. Such an open-ended strategy might be desired by drug discovery groups, for example, who are looking for new approaches to a disease target. As well, in the pending renewal of the Office of Naval Research grant, we have proposed using open-ended searches to identify connections between the disparate fields of gene therapy/gene transfer and viral warfare, in order to anticipate new biotechnologies that have particular potential to impact on the latter field. In the future,

one can even envision a research program that automatically sifts the literature looking for new potential connections between broad disparate areas of inquiry.

Specific Aim 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.

Though a free, public Web site has already been established at http://kiwi.uchicago.edu, which provides a demonstration of Arrowsmith analyses, this site is not yet suitable for widespread use by the scientific community. In this Specific Aim, we will seek to make all of the changes in the Web site that are needed in order to satisfy the needs of the field testers and other public users of the Web site, at least with regard to Medline searches.

Goal 1: Traffic. The size and speed of the Arrowsmith server is currently being upgraded under the auspices of the UIC Department of Psychiatry, to increase the speed and capacity of the server. However, separate programming is needed to enable the server to handle simultaneous multi-user queries and to allow query results to be stored by users for at least several days. **Goal 2: Display and print out of AB_i and B_iC titles**. At present, all such titles are displayed, regardless of their number, and this creates a mild problem in handling sets that consist of scores of titles. In Specific Aim 1 and 2, we are testing whether it is feasible to truncate or subdivide the AB_i and B_iC titles without jeopardizing the query results. However, even if all AB_i and B_iC titles are displayed for inspection on-screen, one would still like to have the option of printing out only those AB_i and B_iC titles that are deemed informative. **Goal 3: Instructional materials**. Although the site is written with clear directions that explain the mechanics of conducting Arrowsmith analyses, the instructions are not directed toward the general community of neuroscientists, many of whom are not well versed in conventional Medline searching, and who may not know what kind of problems are suitable, nor how to construct a query shrewdly. Therefore, adding new step-by-step instructional materials is a high priority. **Goal 4: Compatibility across browsers and platforms**. At present, the Web site is definitely compatible with users who use the Netscape browser from Windows-PC computers, but testing and new programming are likely to be needed to ensure that it is fully accessible across browsers and platforms. **Goal 5: Documenting the software**. When Arrowsmith was first designed and progressively modified by Don Swanson as a private research tool, there was no need to document the software rigorously. During the current grant period, however, the software will become fully documented so that others can modify and adapt it to their own needs in the future.

Instructional materials. The PI and program manager will be responsible for writing these in a clear and self-explanatory fashion. A tutorial has already been written that covers the simplest type of Arrowsmith search (29), but this is insufficient for serious users. Experience gained in teaching the field testers and analyzing their queries will be invaluable for learning what aspects are most likely to represent potential problems, and need special attention. There will be a "short version" for those familiar with Medline searching, and a more complete version that covers the short-course curriculum (see Specific Aim 1). Though the UIC and U of Chicago Web sites will offer similar software features, they may not be identical at any given time; for example, we plan to experiment with different ways of presenting instructional materials and display formats on the UIC site without automatically making the same changes to the U of Chicago site.

Evaluation. Since the field testers will utilize the Web site for most queries, they will able to provide ongoing feedback regarding the layout, instructions, accessibility and compatibility of the site overall. The instructional materials are targeted primarily to public users of the site, whose numbers are currently small but are expected to grow as the site is modified to allow multi-user traffic. At present, the Arrowsmith site is (to our knowledge) linked only to a single neuroscience site, the Alzheimer Research Forum. As soon as the site can handle a significant amount of traffic, and new instructional materials have been added, we will attempt to publicize the site and link it to other sites frequented by neuroscientists. Ultimately, should Arrowsmith prove to

be suitable for routine queries, it would be desirable to link it to the Entrez-PubMed site itself. Group statistics will be monitored to measure overall traffic to the site, and whether users come from universities (.edu) or commercial outfits (.com). Because we wish users to view the Web site as an academic service, and not to regard themselves as research subjects, public users will not be required to fill out questionnaires evaluating the site or its instructional materials (though we will encourage voluntary feedback via e-mail). It is planned to carry out systematic studies of public Web users in Phase II -- after field testers have provided their inputs, after issues of Web site compatibility and traffic capacity have been resolved, and after it is clear that the public users represent a large enough and diverse enough sample for statistical purposes.

Alternative approaches. The present proposal focuses on the feasibility of Arrowsmith as a tool to be used by individual neuroscientists, and emphasizes the use of a public Web site. However, an alternative approach is to establish Arrowsmith as a stand-alone product that is installed on users' own computers. Particularly in the case of large potential users, such as pharmaceutical companies and Defense analysts, a stand-alone product would be desirable in order to ensure confidentiality of search results, to facilitate examination of large files, and to save query results for long periods of time. All of the searches, evaluations and programming improvements proposed here would apply equally to either Web-based or stand-alone versions of Arrowsmith, particularly if the stand-alone software is formulated as a "virtual web site" accessed by a browser.

Specific Aim 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.

A unique feature of Arrowsmith is its ability to bridge distinct fields or disciplines. But just as no one group of neuroscientists may be aware of specific findings bridging these disciplines that can be mined using Arrowsmith, so no one group may be in a position to test, experimentally or clinically, the implicit suggestions that are discerned from an Arrowsmith analysis. To date, we have attempted to alert other scientists of significant hypotheses worth testing by publishing short reports in journals read by biomedical scientists. These reports have had some impact on the scientific community at least insofar as other investigators have cited them, and insofar as they may have stimulated others to test some of the hypotheses (11). However, publishing involves a significant delay, is not specifically targeted, and is not easily evaluated in terms of its heuristic effectiveness – since all of the component findings of Arrowsmith are already implicit in the literature, someone might already have been thinking and planning along those lines independently. For these reasons, we believe that it is worth exploring more direct interventions to facilitate follow-up of Arrowsmith analyses that result in especially interesting suggestions. In particular, we feel that neuroscientists will respond most favorably to direct human contact – a phone call or personal letter – that is initiated under the aegis of the project.

Procedure. When field testers discern implicit suggestions in Arrowsmith searches that fall into Category 3, he/she will decide, together with the subcontract PI, whether it is worth following up within the field tester's group. To facilitate the testing of pilot studies, some funds are requested for laboratory supplies and related expenses. If it appears that the study would benefit from collaborations with other investigators in some other discipline, then they will contact the relevant scientists directly to make such arrangements. Suggestions that fall into Category 4 will be considered by the field tester, contract PI, program manager and PI to decide whether this suggestion is worth attempting to bring together two sets of relevant scientists from the outside. The most relevant scientists are likely to be those who made the literature reports detected by the Arrowsmith search. If these investigators are not in a position to follow up the findings, or are not interested, then other investigators will be contacted, using informatics information to identify those who are active and well-cited in the field. As a unique "carrot" to facilitate the testing of pilot studies by outside investigators, a modest pool of funds from the UIC budget is earmarked to support laboratory supplies and related expenses in these collaborations. At the least, we will be able to learn how outside investigators rate these Arrowsmith

suggestions, and whether they find them novel and significant. At best, new lines of investigation may be stimulated.

Evaluation. This represents a new approach to scientific collaboration that may be regarded as unusual or even naïve in the hyper-competitive, hyper-specialized environment of today. Yet we are looking to (and hoping to bring about) the neuroscience environment of the future, where sharing of data and ideas will become the norm rather than the exception. We will keep records of the number of Category 4 suggestions that arise per Arrowsmith search; our ability to contact relevant investigators; their reaction to these Category 4 suggestions; and whether a collaboration actually occurs as a result of our contacts. In the long term, though not necessarily during the project period, it will also be relevant to learn if the suggestion is experimentally confirmed and whether the new line of investigation ends there or stimulates further questions.

Possible pitfalls and alternative approaches. It is possible that few Arrowsmith searches will suggest fruitful new collaborations (though this has not been our experience to date), or that outside collaborators will be resistant to testing suggestions based on Category 4 Arrowsmith searches, even with the prospect of support for pilot studies. If so, then we may refocus attention on tracking Category 3 searches – these may have a better likelihood of successfully recruiting new collaborators, since one of the field testers' groups would have a direct interest in, and can contribute expertise to, the proposed experiments being described to outside collaborators. Alternatively, as advocated by Kostoff (18), we may invite a small group of investigators from the two complementary fields to attend a small meeting to highlight the opportunities for collaboration, diverting the supply funds to support travel expenses for the investigators instead.

Timetable and Transition to Phase II trials

All Years: Conduct, evaluate and follow-up individual Arrowsmith searches. Analyze search outcomes in the aggregate and improve the implementation of the software in response to feedback. Assess scientific hypotheses arising from Arrowsmith searches, by laboratory testing and by fostering collaborations with other investigators having complementary expertise. Document the Arrowsmith software and all improvements made.
Year 1: Recruit and train personnel; upgrade the Web site to accommodate simultaneous multi-user queries and to ensure cross-platform compatibility. Incorporate MeSH subject headings into the Web site.
Year 2: Write and evaluate instructional materials for the Web site. Evaluate whether MeSH subject headings can enhance standard title word searches. Explore feasibility of incorporating abstract terms.
Year 3-5: Extend the scope of Arrowsmith as warranted by user needs, for example, to include custom databases, bibliographic databases other than Medline, stand-alone software, or new strategies for constructing B-lists.
This Phase I project should generate the following expected products:

- Peer-reviewed publications that analyze different features of Arrowsmith (e.g., summarizing user performance on standard vs. expanded Arrowsmith searches, or investigating MeSH and abstract terms)
- Improved and expanded software capabilities that have been validated by field use;
- A Web site ready for general public use and further evaluation in Phase II;
- A short course and instructional materials that are suitable for training a larger body of neuroscientists at UIC and elsewhere in online searching and Arrowsmith analyses;
- New scientific hypotheses arising from Arrowsmith analyses, together with tests of those hypotheses;
- New interdisciplinary collaborations between groups of investigators.

e. Human Subjects

None.

f. Vertebrate Animals

None.

g. Literature Cited

1. Swanson, D. R., (1986) Undiscovered public knowledge. Library Quarterly 56: 103-118.

2. Swanson, D. R. and Smalheiser, N. R. (1997) An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artificial Intelligence 91: 183-203.

3. Koslow, S.H. and Huerta, M.F. (1997) Neuroinformatics: An Overview of the Human Brain Project. Lawrence Erlbaum Assoc., Mahwah, N.J.

4. Swanson, D. R. (1986) Fish oil, Raynaud's Syndrome, and undiscovered public knowledge. Perspect. Biol. Med. 30: 7-18.

5. Swanson, D.R. (1988) Migraine and magnesium: eleven neglected connections., Perspect. Biol. Med. 31: 526-557.

6. Swanson, D. R. (1990) Somatomedin C and arginine; implicit connections between mutually-isolated literatures, Perspect. Biol. Med. 33: 157-186.

7. Smalheiser, N. R. and Swanson, D. R. (1994) Assessing a gap in the biomedical literature: magnesium deficiency and neurologic disease, Neurosci. Res. Commun. 15: 1-9.

8. Smalheiser, N. R. and Swanson, D. R. (1996) Indomethacin and Alzheimer's disease. Neurology 46: 583.

9. Smalheiser, N. R. and Swanson, D. R. (1996) Linking estrogen to Alzheimer's disease: an informatics approach. Neurology 47: 809-810.

10. Smalheiser, N. R. and Swanson, D. R. (1998) Calcium-independent phospholipase A_2 and schizophrenia. Arch. Gen. Psychiat. 55: 752-753.

11. Swanson, D. R. (1993) Intervening in the life cycles of scientific knowledge. Library Trends 41: 606-631.

12. Gordon, M. D. and Lindsay, R. K. (1996) 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. Info. Sci. 47: 116-128.

13. Gordon, M.D. and Dumais, S. (1998) Using latent semantic indexing for literature based discovery. J. Am. Soc. Info. Sci. 49: 674-685.

14. Small, H. (1999) A passage through science: crossing disciplinary boundaries. Library Trends 48: 72-108.

15. Trybula, W.J. (1997) Data mining and knowledge discovery. Ann. Rev. Info. Sci. Technol. 32: 197-229.

16. Finn, R. (1998) Program uncovers hidden connections in the literature. The Scientist May 11: 12-13.

17. Valdes-Perez, R.E. (1998) Text-based informatics. The Scientist July 6: 10.

18. Kostoff, R.N. (1999) Science and technology innovation. Technovation 19: 593-604.

19. Spasser, M.A. (1997) The enacted fate of undiscovered public knowledge. J. Am. Soc. Info. Sci. 48: 707-717.

20. Valdes-Perez, R.E. (1999) Principles of human-computer collaboration for knowledge discovery in science. Artificial Intelligence 107: 335-346.

21. Ross, B.M., Hudson, C., Erlich, J., Warsh, J.J. and Kish, S.J. (1997) Increased phospholipid breakdown in schizophrenia: evidence for the involvement of a calcium-independent phospholipase A_2. Arch Gen. Psychiat. 54: 487-494.

22. Kuo, C.-F., Cheng, S. and Burgess, J.R. (1995) Deficiency of vitamin E and selenium enhances calcium-independent phospholipase A_2 activity in rat lung and liver. J. Nutrition 125: 1419-1429.

23. McCreadie, R.G., MacDonald, E., Wiles, D., Campbell, G. and Paterson, J.R. (1995) Plasma lipid peroxide and serum vitamin E levels in patients with and without tardive dyskinesia, and in normal subjects. Br. J. Psychiat. 167: 610-617.

24. Ross, B.M. (1998) In reply. Arch Gen. Psychiat. 55: 753.

25. Tanabe, L., Scherf, U., Smith, L.H., Lee, J.K., Hunter, L. and Weinstein, J.N. (1999) MedMiner: an Internet text-mining tool for biomedical informatics: application to gene expression profiling. Biotechniques 27: 1216-1217.

26. Bookstein, A., Klein, S.T. and Raita, T. (1998) Clumping properties of content-bearing words. J. Am. Soc. Info. Sci. 49: 102-114.

27. Chen, J., Martinez, J., Kirchhoff, A., Ng, T.D. and Schatz, B.R. (1998) Alleviating search uncertainty through concept associations: automatic indexing, co-occurrence analysis, and parallel computing. J. Am. Soc. Info. Sci. 49: 206-216.

28. Gordon, M.D. (1999) Using textual information for discovery and innovation. Proc. 62nd annual meeting of ASIS 36: 771-772.

29. Smalheiser, N. R. and Swanson, D. R. (1998) Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. Computer Meth. Prog. Biomed. 57: 149-153.

30. Fayyad, U.M., Piatetsky-Shapiro, G. and Smyth, P. (1996) From data mining to knowledge discovery in databases. AI Magazine 17: 37-54.

31. Qin, J. (1999) Discovering semantic patterns in bibliographically coupled documents. Library Trends 48: 109-132.

32. Kostoff, R.N., Eberhart, H.J. and Toothman, D.R. (1997) Database tomography for information retrieval. J. Information Sci. 23: 301-311.

h. Consortium/Contractual Arrangements

This application involves a primary site (UIC) which coordinates the entire project; a secondary subcontract site (University of Chicago) at which Don Swanson will continue his research on extending the scope of Arrowsmith software; and 3 subcontract field test sites, each an active group conducting research in neuroscience and neuro-informatics. These groups are headed by Drs. Michael Gabriel (Univ. Illinois at Urbana), Maryann Martone (UCSD) and Allan Reiss (Stanford); each of them needs their own budget, since they are responsible for recruiting and supervising field testers, and for following up suggestions that arise from Arrowsmith searches. (It should be noted that Dr. Gabriel works within the University of Illinois system, so that transfer of funds will be internal rather than among separate universities. Face pages, letters of intent, and checklists have been included only from the 3 non-Univ. of Illinois subcontractors.) In addition, Dr. Michael Knable (Stanley Foundation) has offered to host a 4th field test site. However, because of the Stanley Foundation policy of remaining independent of government funding, they prefer not to receive a budget, but instead plan to participate as intellectual collaborators. Letters of collaboration are included from all subcontractors, and from Dr. Knable.

i. Consultants

No individuals are identified as consultants. However, letters affirming their roles in the project are included from Drs. Gabriel, Swanson, Martone and Reiss (subcontract PIs), Dr. Knable (collaborator), and Dr. Flaherty (Head, UIC Dept. of Psychiatry, affirming institutional support).