JESLIB 2013; 2(1): 35-40
doi:10.7191/jeslib.2013.1029

Journal of eScience Librarianship
putting the pieces together: theory and practice

Informationist Support for a Study of the Role of Proteases and Peptides in Cancer Pain

Alisa Surkis,[1] Aileen McCrillis,[1] Richard McGowan,[1] Jeffrey Williams,[1] Brian L. Schmidt,[2] Markus Hardt,[3] Neil Rambo[1]

[1] New York University School of Medicine, New York, NY, USA
[2] New York University College of Dentistry, New York, NY, USA
[3] The Forsyth Institute, Cambridge, MA, USA

Abstract

Two supplements were awarded to the New York University Health Sciences Libraries from the National Library of Medicine's informationist grant program. These supplements funded research support in a number of areas, including data management and bioinformatics, two fields that the library had recently begun to explore. As such, the supplements were of particular value to the library as a testing ground for these newer services.

This paper will discuss a supplement received in support of a grant from the National Institute of Dental and Craniofacial Research (PI: Brian Schmidt) on the role of proteases and peptides in cancer pain. A number of barriers were preventing the research team from maximizing the efficiency and effectiveness of their work. A critical component of the research was to identify which proteins, from among hundreds identified in collected samples, to include in preclinical testing. This selection involved laborious and prohibitively time-consuming manual searching of the literature on protein function. Additionally, the research team encompassed ten investigators working in two different cities, which led to issues around the sharing and tracking of both data and citations.

The supplement outlined three areas in which the informationists would assist the researchers in overcoming these barriers: 1) creating an automated literature searching system for protein function discovery, 2) introducing tools and associated workflows for sharing citations, and 3) introducing tools and workflows for sharing data and specimens.

The Opportunity

The National Library of Medicine request for applications for administrative supplements to fund informationist support of research grants came at an opportune moment for the New York University Health Sciences Libraries (NYUHSL), arriving at the start of a strategic planning process. The supplements provided a unique opportunity to develop test cases in areas in which service development was in the initial phases. NYUHSL received two supplements: one for a project involving a number of issues in data management, described elsewhere in this issue (Hanson et al. 2013), and one for the project described here, which involves data management and bioinformatics support, as well as established services of citation management support and literature searching. In undertaking an analysis of what the NYUHSL model for library services should

Correspondence to Alisa Surkis: alisa.surkis@med.nyu.edu
Keywords: cancer, neoplasms, pain, proteases, peptides, proteins, informationists, data management, citation management, bioinformatics

look like over the next several years -- what services are needed, how to best demonstrate their value, and how they could be supported and scaled -- these opportunities will provide invaluable information.

The Grant

An administrative supplement was received for informationist support for a National Institute of Dental and Craniofacial Research grant entitled "Role of proteases and peptides in cancer pain," that was led by Dr. Brian Schmidt, a surgeon-scientist at the NYU College of Dentistry, along with co-principal investigator Dr. Markus Hardt, a protein chemist at the Boston Biomedical Research Institute. Pain is a primary concern in the cancer patient, but the etiology of cancer pain is unclear, and elucidation of these mechanisms is of great importance, as it may lead to targeted analgesics for cancer patients (Schmidt et al. 2010). The principal investigators' work is focused on identifying and determining the potential pain-mediating role of proteins and peptides within the microenvironment of oral squamous carcinomas. Their discovery approach (Hardt et al. 2011) produces several hundreds of candidate molecules that require screening and validation.

The workflow for identifying these proteins begins with screening and enrolling patients with biopsy-proven oral cancer, having them complete an oral cancer pain questionnaire, and then, at the time of surgical resection, performing microdialysis on the cancer and on an anatomically matched contralateral normal site. The microdialysis samples are then sent to the Hardt laboratory for identification and characterization of the proteases and peptides, using mass spectrometry and activity based chemistry. High-probability molecules (i.e. those molecules that are likely contributing to cancer pain) are identified by the researchers through review of the candidate list and identifying candidates based on their personal fund of knowledge and on the literature. Once these candidates are chosen, they are either isolated from the sample or synthesized, and then shipped back to the Schmidt lab where nociceptive (i.e. pain-producing) capacity is determined in a preclinical model.

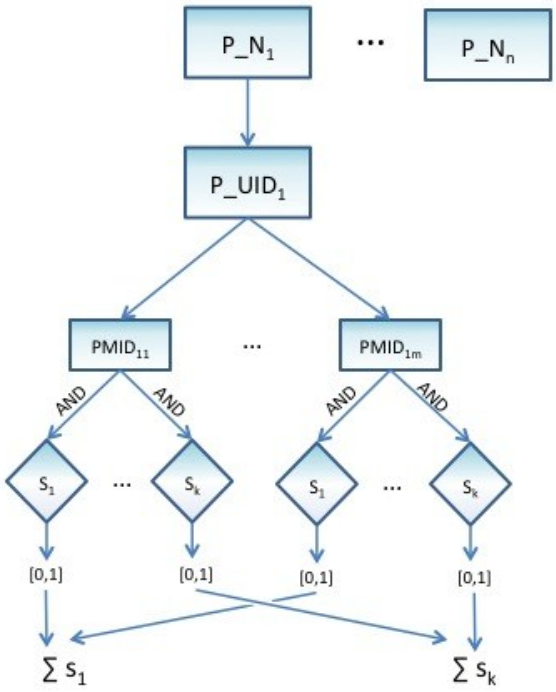
The Problems

Contained within the workflow above are several points where significant barriers were encountered to achieving the aims of the parent grant. These barriers were the cause of inefficiency in the research process, the introduction of data errors, and a lack of maximum effectiveness.

A thorough review of the literature is needed to determine which of the hundreds of proteases and peptides identified have known associations with cancer pain, and which have not yet been characterized as to their function. This review is crucial to the investigation, but the method of literature searching used was keyword searching in PubMed and Google Scholar, a cumbersome and time-consuming approach, allowing only a very limited amount of searching to be done. Another problem is related to citation management. Different investigators used different methods or tools for managing their citations. Information sharing was cumbersome, and the process of including citations and bibliographies during manuscript preparation was unnecessarily inefficient.

Finally, problems were arising in the transfer of data and specimens between the two labs. Peptides and proteases are characterized in a number of different samples, and for each sample a very large file is created containing the data resulting from the analysis. Samples may be manipulated further, sent back and forth between sites, with additional information about what is contained in the sample being added to these files. Following identification of the proteases or peptides within the samples, the samples are fractionated or the molecules of interest are synthesized. The fractions or synthesized molecules are then sent between laborato-

Figure 1: Workflow from protein name (P_N) through protein identifier (P_UID) to relevant papers (PMID), which were then searched for each of the relevant search terms (S) to establish a measure of known associations of the protein with the concept (∑S).



ries. This process creates a number of data management challenges: a complicated workflow, research being conducted at geographically disparate locations, and large data files. This situation led to time-consuming and costly data errors.

The Solutions

The first specific aim of the supplement is to address issues encountered in searching the literature for known roles of identified proteins. It is critical that some automation be introduced into the searching process for literature about these hundreds of proteins. The solution proposed was to programmatically access the linked NCBI databases through the use of the Entrez programming utilities (NCBI 2010). This would allow the databases to be automatically queried for each of the many proteins in order to locate linked PubMed articles, then to and identify those articles within which the search concepts of interest were referenced. Working in the Matlab programming environment, an output matrix would be produced to consolidate results, rank them by relevance and strength of known associations, and identify those proteins without known associations. This customized, synthesized output would immediately enhance the discovery of relevant scientific literature, and biological data. Additional support would be provided in developing optimal search query terminology and syntax, and searching a wider array of resources, such as Biobase PROTEOME

and ISI Web of Knowledge.

The second aim of the supplement is to provide assistance in the organization, management, and sharing of bibliographic information. Use of a common bibliographic management tool and establishment of a shared library would allow greater coordination among the research team, increase the efficiency of information transfer, and simplify the process of integrating references and bibliographies into publications. The informationists propose to guide the development and implementation of a streamlined workflow to manage references and other information to meet the specific needs of the research team.

The third aim is to develop a web-based method for data management and specimen cataloging. The informationists proposed to perform an assessment of data management needs in order to best support the research team in producing and implementing a data management plan to improve data quality and consistency. Issues to be addressed included efficient data structuring, development of data workflows, thorough and transparent documentation, and proper use of metadata, including use of established standards when available. An analysis of existing tools will be undertaken to assist in establishing an online collaborative space to allow researchers at both sites to access and modify data files, while keeping those files secure, attaching necessary documentation, and introducing some controls over versioning.

Progress to Date

Work to date has focused on the identification of known protein functions from the literature. Informationists assisted in identifying an initial set of search terms and wrote a Matlab program to call the NCBI programming utilities and process the output of those calls. The program (Figure 1) first uses protein names to retrieve protein unique identifiers, then uses those identifiers to retrieve linked articles from PubMed. It then searches each article for each of the search queries, and finally, produces an output matrix containing the overall number of publications identified for each query for each protein, and the PubMed IDs of each of those publications. A further modification is underway to filter the papers by the Impact Factor of the journals in which they were published. Only a limited number of proteins can be included in preclinical testing, so those for which the pain-related functions were published in higher impact journals are assumed to be more promising candidates.

Initial analysis of citation management tools has narrowed the choices to Endnote Web and Mendeley. The next step in this process is for the informationists to meet with the research team to present the advantages and disadvantages of each of these options, and establish which tool will best meet their needs.

Work has not yet begun on the data management component of the supplement, but initial conversations have indicated REDCap as a leading contender for a tool to use for this process. It is supported by the NYU Langone Medical Center, and is currently being evaluated for use in another of the PI's projects. The next step in this aspect of the project is for the informationists to further familiarize themselves with REDCap by making use of available tutorials and manuals, and to evaluate whether REDCap will meet the needs of the researchers.

Evaluation Plan

The evaluation plan for informationist support involves the use of formative and summative assessment techniques to evaluate the success and impact of the informationists' work in support of the strategic aims. Formative assessments – both formal and informal – will dictate changes in the informationists' approach. Summative assessment will be conducted via quarterly reports to the funding agency that will document pro-

gress on addressing key research team problems in 1) literature and biological database searching; 2) organization, management and sharing of bibliographic information; and 3) data management and specimen cataloging.

A member of the NYUHSL's Administrative Team will conduct an interview with both Principal Investigators to solicit and document their assessment of the value and impact of the informationists to the research team's efforts. The interview will address issues such as whether previously undetected literature on proteins had been located, whether the PIs were spending less of their time identifying literature, and whether the research team was experiencing fewer data errors.

Broader Outcomes and Scalability

In addition to the goal of facilitating the achievement of the aims of the parent grant, an additional desired long-term outcome is to contribute to improvement of research efficiency throughout the institution, via the extension of library services. An important component of achieving the latter goal involves raising awareness of the value that informationists can bring to research projects, through marketing and presentations outside of the library community. Initial efforts in that direction include a press release by the NYU College of Dentistry (Anon. 2012) and a poster submission on the protein function discovery workflow to the 2013 American Medical Informatics Society Joint Summits on Translational Science.

A critical issue when considering how informationist services could fit into the overall NYUHSL service model is that of scalability. The time commitment for both supplements is substantial, and careful consideration would have to be given to scaling beyond the two informationist projects currently underway. One option is not to scale up, but rather to maintain this as a specialty service, with the library seeking out or accepting such opportunities only very sparingly. Another possibility is that, whether through further supplements or informationists being written directly into parent grants, support is found to expand the pool of available informationists. A third option is to focus on service expansion in areas where it will be possible to re-use work from these initial projects, thus maximizing the impact for the time input.

One further point, is that this project calls upon a number of skill sets across the team of informationists, ranging from skills that are well within the comfort zone of the informationists, such as citation management support, to those more recently acquired, such as the use of the linked NCBI databases and Entrez programming utilities. Other skills, such as the use of a web-based data management tool, will require additional training. Expansion and solidification of the types of skills needed for these projects should positively impact the scalability of these services.

References

Anon. 2012. "NYU College of Dentistry's Dr. Brian L. Schmidt and Boston Biomedical Research Institute's Dr. Markus Hardt Share NIH Grant to Collaborate with Informationists." *NYU News*. http://www.nyu.edu/about/news-publications/news/2012/11/29/nyu-college-of-dentistrys-dr-brian-l-schmidt-and-boston-biomedical-research-institutes-dr-markus-hardt-share-nih-grant-to-collaborate-with-informationists.html.

Hanson, Karen, Theodora Bakker, Mario Svirsky, Arlene Neuman, and Neil Rambo. "Informationist Role in Clinical Data Management." *Journal of eScience Librarianship* 2, no. 1 (2013): 25-29, http://dx.doi.org/10.7191/jeslib.2013.1030

Hardt, Markus, David K. Lam, John C. Dolan, and Brian L Schmidt. "Surveying Proteolytic Processes in Human Cancer Microenvironments by Microdialysis and Activity-

based Mass Spectrometry." *Proteomics Clinical Applications* 5, no. 11-12 (2011): 636-643, http://dx.doi.org/10.1002/prca.201100015

National Center for Biotechnology Information. *Entrez Programming Utilities Help.* Bethesda (MD): National Center for Biotechnology Information (US), 2010. http://www.ncbi.nlm.nih.gov/books/NBK25501/

Schmidt, Brian L., Darryl T. Hamamoto, Donald A. Simone, and George L. Wilcox. "Mechanism of Cancer Pain." *Molecular Interventions* 10, no. 3 (2010): 164-178, http://dx.doi.org/10.1124/mi.10.3.7

Acknowledgements

This work was supported by NLM grant R01DE019796-03S1.

Disclosure: The authors report no conflicts of interest.

ISSN 2161-3974