Information Research, Volume 6 No. 2 January 2001
Facilitating access to and use of bioinformatics resources,by **Joan C. Bartlett**
Location: http://InformationR.net/ir/6-2/ws1.html

Facilitating access to and use of bioinformatics resources

Joan C. Bartlett
Faculty of Information Studies
University of Toronto

Introduction

As the Human Genome Project nears completion, a vast and ever-expanding body of bioinformatics resources is being developed. This includes databases of biological information, and software tools that manipulate and analyze the data. Bioinformatics resources have significantly affected the conduct of biomedical research. However, very little research has been conducted into how scientists use these resources. This paper presents proposed dissertation research[1] into one approach to facilitating access to bioinformatics resources.

Background

The data generated through molecular biology research has led to the development of a wide range of molecular biology databases, non-bibliographic databases of information such as DNA sequences (e.g., GenBank), protein sequences (e.g., Swiss-Prot), and genomic mapping information (e.g., Genome Database). The development, management and use of these databases is one component of the field of bioinformatics, "the computer-assisted data management discipline that helps us gather, analyze and represent this information" (Persidis, 1999: 828).

The Human Genome Project began in 1990, with the goal of sequencing the entire human genome, localizing the estimated 100 000 genes, and creating a genetic and physical map of the genome (National Human Genome Research Institute, 2000). The completion of a 'working draft' of the human genome sequence was announced June 26, 2000, and one-third of the genome is anticipated to be completely sequenced by 2001 (Wadman, 1999). The current goal is to have the complete sequence by 2002, two years ahead of the schedule set in 1990 (Baxevanis, 2000: 1).

The latest annual database issue of the journal Nucleic Acids Research listed over 200 molecular biology databases (Baxevanis, 2000). The information in these databases includes sequences (DNA, RNA, and protein), gene structure, maps, mutations and genomes, among others. The GenBank database alone grows an average of 2 million base pairs per day (Baxevanis, 2000). The National Library of Medicine ENTREZ system, the main search interface to both bibliographic (e.g., Medline, HealthSTAR) and molecular biology databases (e.g., GenBank), is queried 50 000 times per day (Persidis, 1999).

As the number and size of molecular biology databases grow and their use increases, concerns have been raised about the lack of expertise in the area of bioinformatics (Collins *et al.*, 1998; MacLean & Miles, 1999). Scientists are overwhelmed by the range and volume of bioinformatics resources available to them, and frequently lack the expertise to use them (Yarfitz & Ketchell, 2000). Anecdotal evidence shows that it is not uncommon for scientists to spend a long period of time conducting wet lab research to obtain results that could also be obtained through the use of bioinformatics resources. The use of these resources can lead to savings of both time and money for the researcher.

To date, there has been little research into the use of these essential research resources. However, as bioinformatics resources continue to expand and develop in complexity and number, it is essential to understand the users, what information they need from the resources, how they search for information, and how they use that information. The results of research in these areas could be used for many purposes. They can inform the design and development of bioinformatics resources in order to make them responsive to the needs of the users. They can be used as a basis for developing training and education programs, so that scientists may become better informed and more proficient in the use of bioinformatics resources. Research could also lead to the development of tools to guide and assist scientists when they are using these resources.

Much of the past research into the information behaviour of life scientists and molecular biologists has focused on the use of traditional, bibliographic resources (Bayer & Jahoda, 1981; Curtis, Weller, & Hurd, 1993, 1997; Dillon, 1981; French, 1990; Garvey, Tomita, & Woolf, 1974; Palmer, 1991a, 1991b; Rolinson, Meadows, & Smith, 1995; Skelton, 1973). Little research has addressed the use of non-bibliographic bioinformatics resources. Studies of molecular biologists found that, as early as 1991, molecular sequence databases were seen as having an increasingly important role, with the need to either become proficient with their use, or be left behind (Grefsheim *et al.*, 1991). At the same time, there was a call for help in managing the expanding array of information available (Grefsheim *et al.*, 1991). Almost 10 years later, a study found that 70% of molecular biologists surveyed were using molecular sequence databases on a weekly or monthly basis (Yarfitz & Ketchell, 2000). While personal contacts with colleagues was the primary source of information about these resources, there was considerable interest in bioinformatics consultation services, classes, and other services to provide assistance in the use of bioinformatics resources (Yarfitz & Ketchell, 2000).

Research objectives

The intent of my research is twofold: a) to assess the patterns of use of bioinformatics resources (both molecular biology databases and related analytical software tools) by expert users, and b) to create and test a tool, based on a predictive model, that will facilitate the access and use of bioinformatics resources by scientists. Among the challenges facing scientists is the appropriate choice and use of bioinformatics resources. This research will analyze how experts use them, to capture patterns of use. That is, for scientists engaged in a particular type of research, are there patterns or common features to their use of bioinformatics resources? From these patterns, I will build a predictive model for the use of bioinformatics resources for a particular scientific task, and use the model to develop a tool to assist scientists in their selection and use.

There are two specific research objectives:

1. To identify and assess the patterns of use of bioinformatics resources by experts, so that a predictive model can be developed.

2. To create and test a computer-based tool, based on the predictive model, that can assist novices to access and use bioinformatics resources.

Research design

Phase I

Objective: To identify and assess the patterns of use of bioinformatics resources

The objective of the first phase is to identify and assess the patterns of use of bioinformatics resources by expert users, scientists who are proficient in their use. The scientists will all be involved in similar types of research, so that the tasks for which they require information will be similar.

Twenty scientists will be interviewed using a semi-structured interview script. This will allow the scientists to recount their experiences in their own words, discussing what they consider to be important, while still allowing the researcher to ensure that key points are covered. The focus of the interview will be to understand the scientific task prompting the use of bioinformatics resources, and the process by which they were used. The initial interview question will ask the scientist to describe the steps they would follow, and the databases and tools they would use to complete a scientific task. Probe questions will be used to ensure that essential points are covered. These will include:

- What sequence of events was followed?
- What databases and tools were used?
- What triggers this process?
- What information do you have at the start of the process?
- Does that information have to be prepared in any way?
- What information comes out of the process at the end?
- What happens after this?
- How do you know when you are finished?

Additional questions will explore the amount of variation in the process as described by the scientist, in order to understand if the process could be generalized to other situations. It is also important to understand the key decision points in the process, why particular resources were chosen, and the purpose of each type of analysis.

The data from the interviews will be analyzed to capture patterns in the use of bioinformatics resources. That is, for a given scientific task, are the bioinformatics resources accessed and utilized in a consistent pattern. From these patterns, a predictive model will be developed. This will be an iterative process, with the model being developed as the data is analyzed, and tested and refined with each new set of data.

In addition to understanding the processes followed by the scientists, it will also be necessary to understand the various resources used. Analysis of the resources will involve a taxonomic classification. It is important to understand the purpose and functions of the various databases and analytical tools, in terms of how each one is used, the capabilities of each, and the similarities and differences. It will be particularly important to understand which resources have similar or overlapping functions, and which are different.

The interview data will also be analyzed to understand the reasons behind the processes described. It will be important to understand not only what resources were used, but how and why. The qualitative analysis will enhance the quantitative focus that will lead to the development of the model.

Phase II

Objective: To develop and test a tool to facilitate access to and use of bioinformatics resources

The second phase will involve the creation and evaluation of a tool to assist scientists to access and use bioinformatics resources. The tool will be developed based on the model from Phase I, and could take several forms. Among these are a search guide, a decision support tool, or training materials. The findings from the interviews will help determine the most appropriate form of the tool. The exact form of the tool, or what it will cover will not be known until the data from the first phase has been analyzed.

Once developed, the tool will be tested with a group of scientists who are not expert in bioinformatics, to determine whether it can facilitate the access to and use of bioinformatics resources. The testing will evaluate both the model on which the tool was based, and the tool itself. However, the emphasis will be on testing and refining the predictive value of the model and its use as the basis of a tool, rather than on the usability of the tool. Test participants will be assigned a standardized scientific task (such as the generation of a restriction map, or the characterization of a novel sequence), for which bioinformatics resources could provide a solution. They will be asked to solve the problem, and will be randomly assigned to either use the tool, or not. Among the factors to be considered in the evaluation are efficiency, effectiveness and user satisfaction.

Outcomes

This research will determine whether the usage patterns of bioinformatics experts are sufficiently predictive to assist other (non-bioinformatics expert) scientists to access and use bioinformatics resources. It is anticipated that there will be patterns that can be followed to complete a scientific task.

An simple example of such a task is the determination of how to remove a specific segment of DNA from a bacterial chromosome. First, the entire sequence of the chromosome must be identified. This information is available from publicly accessible sequence databases. Then, the retrieved sequence must be analyzed by a computer program which searches for and identifies the location of the specific points on the DNA which will be cut by specialized enzymes. The output of the analysis is a map of all of the cut sites on the chromosome. The researcher would then identify those sites which flank the segment of interest that they wish to remove.

This example, while describing a fairly simple scientific task, nonetheless demonstrates the type of patterns which may emerge from the usage patterns of experts. The database and analytical tool used may vary among scientists, and the choice may be influenced by several factors, including the specific details of the task (for example, the organism being studied), the resources available, and personal preference. However, the class of resource used would be predictable.

It is expected that this dissertation research will lead to the development of a tool to assist scientists in the application of bioinformatics resources to one particular task. This could provide a solution to the problem of scientists being overwhelmed by the range of resources available to them, and assist them in navigating among them. This research will also investigate the feasibility of using a predictive model, developed by studying expert users, to create a decision support tool, a search guide or training materials. If this is found to be effective, then the approach can be applied to other scientific tasks for which bioinformatics resources are used. The approach could also be applied to other disciplines with comparable information resources.

Note

[1] Dissertation Committee:
Elaine Toms (Supervisor), Associate Professor, Faculty of Information Studies, University of Toronto; Joan Cherry, Professor and Associate Dean, Faculty of Information Studies, University of Toronto; Jamie Cuticchia, Head and Senior Bioinformatics Scientist, Bioinformatics Supercomputing Centre, Hospital for Sick Children, Toronto.

References

- Baxevanis, A. D. (2000). "The Molecular Biology Database Collection: an online compilation of relevant database resources." *Nucleic Acids Research*, **28**(1), 1-7.
- Bayer, A., & Jahoda, G. (1981). "Effects of online bibliographic searching on scientists' information style." *On-Line Review*, **5**, 323-33.
- Collins, F. S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R., & Walters, L. (1998). "New goals for the U.S. Human Genome Project: 1998-2003." *Science*, **282**(5389), 682-9.
- Curtis, K. L., Weller, A. C., & Hurd, J. M. (1993). "Information-seeking behaviour: a survey of health sciences faculty use of indexes and databases." *Bulletin of the Medical Library Association*, **81**(4), 383-392.
- Curtis, K. L., Weller, A. C., & Hurd, J. M. (1997). "Information-seeking behavior of health sciences faculty: the impact of new information technologies." *Bulletin of the Medical Library Association*, **85**(4), 402-10.
- Dillon, M. (1981). "Serving the information needs of scientific research." *Special Libraries*, **72**, 215-23.
- French, B. A. (1990). "User needs and library services in agricultural sciences." *Library Trends*, **38**, 415-41.
- Garvey, W. D., Tomita, K., & Woolf, P. (1974). "The dynamic scientific information user." *Information Storage and Retrieval*, **10**, 115-31.
- Grefsheim, S., Franklin, J., & Cunningham, D. (1991). "Biotechnology awareness study, part 1: where scientists get their information." *Bulletin of the Medical Library Association*, **79**(1), 36-44.
- MacLean, M., & Miles, C. (1999). "Swift action needed to close the skills gap in bioinformatics." *Nature*, **401**, 10.
- National Human Genome Research Institute. (2000). "Human Genome Project Goals: 1998-2003." Bethesda, MD: Retrieved March 24, 2000 from the World Wide Web: http://www.nhgri.nih.gov/HGP/.
- Palmer, J. (1991a). "Scientists and information: I. Using cluster analysis to identify information style." *Journal of Documentation*, **47**(2), 105-29.
- Palmer, J. (1991b). "Scientists and information: II. Personal factors in information behaviour." *Journal of Documentation*, **47**(3), 254-75.
- Persidis, A. (1999). "Bioinformatics." *Nature Biotechnology*, **17**, 828-30.

Rolinson, J., Meadows, A. J., & Smith, H. (1995). "Use of information technology by biological researchers." *Journal of Information Science*, **21**(2), 133-9.

- Skelton, B. (1973). "Scientists and social scientists as information users: a comparison of results of science user studies with the investigation into information requirements of the social sciences." *Journal of Librarianship*, **5**(2), 138-56.
- Wadman, M. (1999). "Human Genome Project aims to finish 'working draft' next year." *Nature*, **398**, 177.
- Yarfitz, S., & Ketchell, D. S. (2000). "A library-based bioinformatics services program." *Bulletin of the Medical Library Association*, **88**(1), 36-48.

Information Research, Volume 6 No. 2 January 2001
***Facilitating access to and use of bioinformatics resources*,by Joan C. Bartlett**
Location: http://InformationR.net/ir/6-2/ws1.html

Last updated: 5th January 2001
Printer-friendly version