Understanding How Experts Use Bioinformatics Resources

Joan C. Bartlett
Faculty of Information Studies
University of Toronto
bartlett@fis.utoronto.ca

ABSTRACT

This paper reports the preliminary findings of a study to understand how experts use bioinformatics resources. These are complex databases of gene or protein sequences, and software tools that manipulate and analyze the data. The experts were seen to follow distinct patterns of resource use, which are quite different from traditional patterns of information-seeking behaviour.

RÉSUMÉ

Cette communication présente les conclusions préliminaires d'une étude qui regardera comment les experts utilisent des ressources bioinformatiques. Cela sont des bases de données des séquences génétiques et de protéines, et des outils de logiciels qui manipulent et analysent les données. On a observé les experts suivant des schémas distincts de l'emploi des ressources qui sont tellement différentes des schémas traditionnels du comportement de la réclame informatique.

INTRODUCTION

Bioinformatics resources include over 600 non-bibliographic databases of biological information such as gene or protein sequences, and software tools that manipulate and analyze the data. These resources are dynamic, complex, diverse, very large, and non-standardized. The ability to access and utilize the information contained in bioinformatics resources will be critical to all areas of future biomedical research, impacting on areas such as disease treatment and prevention, drug development and agriculture. However, the vast range of bioinformatics databases and software poses a significant hurdle to accessing the information, as it is often difficult to know where to begin to find the answer to a specific problem.

This paper presents research focused on understanding how experts use bioinformatics resources. The goal is to understand the patterns of resource use of bioinformatics experts, and to relate the patterns to established models of information behaviour. We report here on the preliminary findings based on interviews with five bioinformatics experts.

BACKGROUND

Since 1953, when the structure of the DNA (deoxyribonucleic acid) molecule was discovered, scientists have sought to unlock the secrets held within the genetic code. DNA is the molecule that carries the blueprint of life. Its familiar double-helix structure (Figure 1) carries the information required to create every living organism. This information is encoded along the strands of the DNA double-helix, with the code written in a four-letter alphabet (Figure 2).

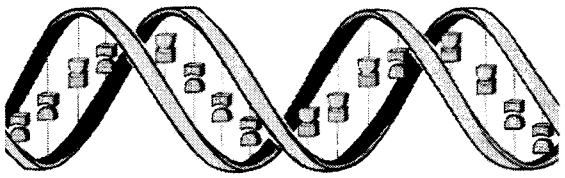

Figure 1. DNA double helix

```
ccacaattacgtcctctcttc
||||||||||||||||||||
ggtgttaatgcaggagagaag
```

Figure 2. Representation of DNA code

The DNA housed in each cell of an organism provides the cell with the blueprint or instructions to build proteins. Each gene is responsible for producing one protein. Ultimately, the sequence and structure of the protein determine its biological function.

The entire human genetic sequence (genome) contains approximately 3 billion DNA letters (Baltimore 2001). As a result of the Human Genome Project, an international initiative to sequence and map the entire human genome, most of the sequence is now known. The completion of a "working draft" of the human genome sequence was announced June 26, 1999 (Wadman 1999). However, the key to unlocking the secrets within the genome is to understand the function of the protein that each gene encodes. This is one of the next great challenges facing biomedical research.

The data generated by initiatives such as the Human Genome Project has led to the development of a wide range of bioinformatics resources, 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 information in these bioinformatics resources includes sequences (DNA, RNA, and protein), gene structure, maps, mutations and genomes, among

others. The GenBank database alone grows an average of 2 million letters of sequence per day (Baxevanis 2000). The National Library of Medicine ENTREZ system, the main search interface to both bibliographic (e.g., Medline, HealthSTAR) and bioinformatics databases (e.g., GenBank), is queried 50 000 times per day (Persidis 1999). Figure 3 presents a portion of a GenBank record, showing a stretch of gene sequence. This type of data is at the heart of most bioinformatics resources.

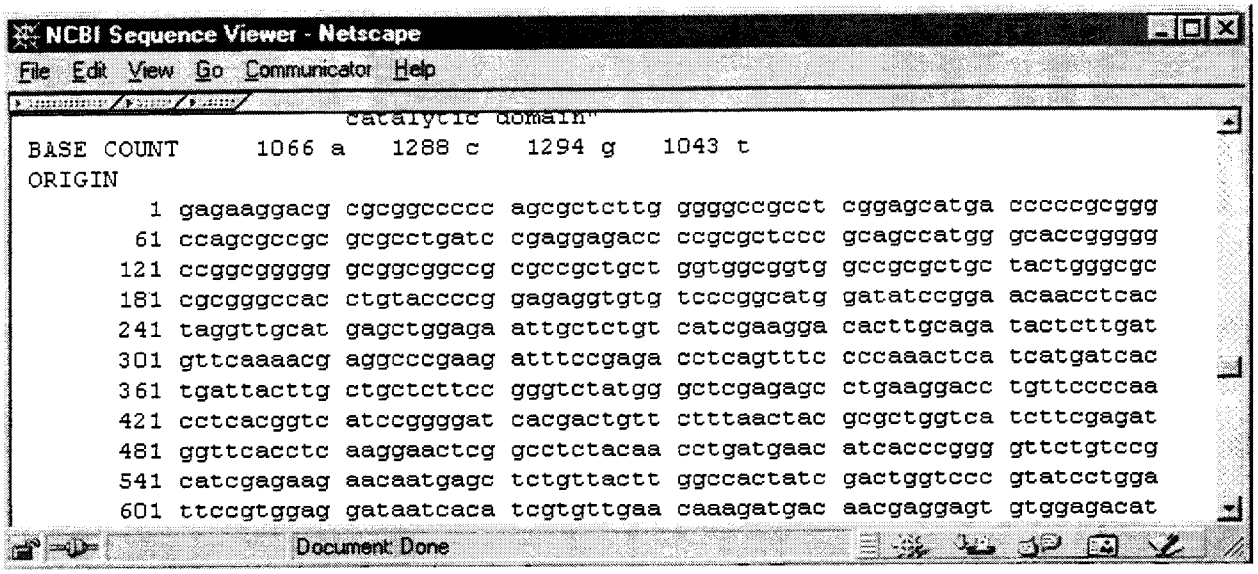

Figure 3: Portion of a GenBank record. This shows part of the human insulin receptor gene sequence. The full sequence is over 4500 characters long. The spaces and numbers are not part of the sequence, but are added to make the data more readable.

As the number and size of bioinformatics resources grow and their use increases, concerns have been raised about the lack of expertise in the area of bioinformatics (Collins et al. 1998; MacLean and 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 and Ketchell 2000). Anecdotal evidence shows that it is not uncommon for scientists to spend a long period of time conducting laboratory research to obtain results that could also be obtained through the use of bioinformatics resources. The use of bioinformatics resources can lead to savings of both time and money for the researcher.

To date, there has been little research or effort to develop tools that facilitate the access to and 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 databases, how they search for information, and how they use that information.

STUDIES AND MODELS OF INFORMATION BEHAVIOUR

Much of the past research into the information behaviour of life scientists and scientists in general has focused on the use of traditional, bibliographic resources (Bayer and Jahoda 1981; Curtis, Weller, and Hurd 1993; Curtis, Weller, and Hurd 1997; Dillon 1981; French 1990; Garvey, Tomita, and Woolf 1974; Palmer 1991a; Palmer 1991b; Rolinson, Meadows, and 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, Franklin, and Cunningham 1991). At the same time, there was a call for help in managing the expanding array of information available (Grefsheim, Franklin, and Cunningham 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 and 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 and Ketchell 2000).

Likewise, models of information seeking behaviour have been developed around conventional types of information queries and sources. Ellis (1989) identified six categories in the information seeking process: *starting, chaining*, *browsing*, *differentiating*, *monitoring*, and *extracting*. In addition to actually obtaining information from a source (extracting) the categories include activities to identify available resources (starting), and to determine what information is available from various resources and which is the best source (browsing, differentiating).

Kuhlthau (1991) proposed a model that described not only stages of the information seeking process, but also the cognitive and affective aspects of the process. Her model has six stages: *initiation*, *selection*, *exploration*, *formulation*, *collection* and *presentation*. The early stages of the model involve the process of deciding on the topic of the search, exploring the available resources both to determine what types of information are available, as well as to consider possible foci of the search. It is not until the later stages of the process that the search topic becomes precise, and search becomes focused.

While the past research into information behaviour has led to fairly consistent findings, these studies have focused on the use of traditional, text-based information

resources. It is not clear whether the findings are applicable to the use of complex, specialized, non-textual resources such as bioinformatics resources.

RESEARCH OBJECTIVES

The intent of this research is to understand how experts use bioinformatics resources to conduct a functional analysis of a gene, and to compare their behaviour to existing models of information behaviour. This work is part of a larger to study to model the information behaviour of bioinformatics experts, and to ultimately apply that model to the development of a tool to facilitate access to bioinformatics resources for bench scientists.

We focused in depth on one particular problem as the foundation for the study: how to characterize and understand the function of a novel gene. That is, given sequence data (nucleic acid or amino acid) for an unknown gene, what information about the characteristics or function of that gene can be obtained from an analysis of the sequence data?

One example of why it is important to conduct a functional analysis of novel sequences is in the area of drug development. A researcher interested in developing a new antibiotic would be interested in finding a gene within the bacteria which could then be targeted by the new drug. Faced with over 1000 bacterial genes with unknown function, it would be extremely valuable to determine which of those genes are involved in processes that could be targeted by the new drug. Several conditions would be established that a gene of interest would have to meet. These could include: involvement in an essential process in the bacteria (so that disruption of the process by the drug would destroy the bacteria), that the gene should not be present in humans (so that the drug would not have undesirable side effects on the patient), and that the protein produced by the gene should be physically located in the cell such that the drug could reach it. By using a variety of bioinformatics tools, it is possible to isolate less than 100 genes (from the over 1000 uncharacterized genes) that meet the conditions of interest. This is a much more manageable number of genes to study in a laboratory setting.

METHODOLOGY

The research involves interviews with bioinformatics experts, aimed at capturing and understanding their expertise.

Participants

The five participants for the interviews were bioinformatics experts. Two were bioinformatics consultants, two were research scientists (with a specialization in bioinformatics) and one was a principal investigator (also with a specialization in bioinformatics). Three people held a Masters degree, and the other two had received a Ph.D. One person had used bioinformatics resources for 1-2 years, three had used them for 3-5 years, and the fifth person had used them for over 5 years. All reported being very confident in their ability to use bioinformatics resources. There were three women and two men, three were in the age range of 26-35, one in the 36-45 range, and one in the 46-55 range.

Participants were intentionally selected from three different research groups, one government, one academic, and one private sector. It was important that the participants be from different groups, so as to minimize the bias towards one particular group's "house style".

The participants were identified through a key informant, a member of the research team who was familiar with the research centres and their personnel.

Interviews

Data were collected using semi-structured interviews with the bioinformatics experts. The semi-structured format allowed the experts to recount their experiences in their own words, discussing what they considered to be important, while still allowing us to ensure that key points were covered.

The participants were given the scenario of having a novel sequence, and asked how they would go about characterizing the sequence, and trying to understand the function of the gene. Following a task analysis approach (Hackos and Redish 1998), the initial question asked the expert to describe the steps they would follow, and the databases and tools they would use to complete the task. Probe questions were used to ensure that essential points were covered. These included questions about the sequence of events that was followed, the databases and tools used, and the information obtained at the end.

Additional questions explored the amount of variation in the process as described by the expert, in order to understand if the process could be generalized to other situations. They were also asked what points of the process they consider to be rate-limiting, and what parts they believe could be automated. Finally, it was important to understand the key decision points in the process, why particular

resources were chosen, and the purpose of each type of analysis. The interviews were audio-recorded, and then transcribed.

In addition to the interview questions, participants were asked to complete a brief background survey. This was intended to obtain demographic information about the participants, as well as some background information about their use of bioinformatics tools.

RESULTS

The preliminary results indicate that experts follow a set procedure of resource use when trying to characterize a novel gene. The procedures may vary depending upon the situation or individual, but, given a similar problem, an expert would follow a consistent routine to arrive at the answer. Thus far, two distinct but related procedures have emerged from the data.

For the task of understanding the function of a novel gene, as many as twenty or thirty different resources can be used in a multistep (6-10 steps) process. The majority of analyses involve comparing the novel sequence against a database of characterized sequences, looking for similar sequence patterns. If a similarity is found, and the pattern is known to be associated with a particular function, then this provides an indication of the possible function of the novel gene.

Central to each of the observed procedures is a building blocks approach to solving the problem of identifying the function of the novel gene. The entire process involves many different types of analyses and resources. However, at each step, only one aspect of the problem is addressed. A specific type of analysis is conducted, with the intent of achieving a specific goal. The results add one more piece to the accumulating body of information. As they are collected, the results from each step are analyzed in an iterative process, with each new piece being compared to previous information, to see if the pieces form a consistent, coherent picture.

In many cases, one step is repeated using multiple tools. Since each tool is based on a unique data set and algorithms, it is possible to get different results, even when conducting similar analyses on the same piece of data. Typically, participants reported repeating an analysis three times, using three different tools. If the analyses all gave a consistent result, then there was a high level of confidence in the finding.

It is also important to look for consistent results at different steps in the process. Each step of the analysis adds to the big picture. So, as the result of each new

analysis is obtained, it must be compared back against the findings thus far, to determine if it is consistent with what has already been found. In cases of inconsistency, one approach is to repeat one or more steps with additional tools.

One point covered in the interviews was what could be done to improve the procedures described by the participants. The most common response referred to the inconsistencies among the variety of bioinformatics resources. Many of the tools used were developed by different research groups or organizations. As such, each had a different structure, a different interface, a different set of parameters that could be specified, and different output. This meant that the input data had to be formatted separately for entry into each tool, and the particularities of each tool had to be learned and understood. The experts' skill was in knowing which tools to use for a particular analysis, how to use them, and how to interpret the results.

An interesting finding was that although the analysis of the sequence with bioinformatics resources can give an indication of the function of the gene, there is still the need to go to the laboratory to verify the findings experimentally. Participants were unlikely to make a conclusion based solely on the evidence from bioinformatics resources, and were sceptical of findings that were reported without experimental confirmation. However, they saw the value of bioinformatics analysis as a means of identifying the most likely candidates for experimental analysis. Faced with a large number of genes that might be of interest (e.g., as the target of a new drug), and the high investment of both time and money to analyze all of the sequences, the ability to quickly and inexpensively identify the most likely candidates is a tremendous advantage. In addition to the immediate savings in reducing the amount of experimental work, this can also lead to a new drug being brought to market more quickly. In this way, bioinformatics analysis was an integral step that fit alongside laboratory analysis. The use of bioinformatics resources can be seen as an extension of traditional laboratory techniques, providing similar types of primary data.

It is interesting to contrast the information seeking patterns of expert users of bioinformatics resources with the traditional information seeking models. The experts have a very focused, directed pattern of information seeking. They have specific tasks that they wish to accomplish in the analysis, with preferred resources to use at each step. There was not any browsing or exploration of the resources to determine which ones to select. Likewise, there was no exploration of the purpose of the search. Instead, the search was focused from the beginning on a specific outcome. Another interesting facet of the process was the stepwise approach. Rather than trying to accomplish everything in one step, the experts added specific pieces of information, one at a time, to the larger picture. They recognized that one source was not sufficient to provide all of the information required, and that

multiple, specialized resources were required.

The procedures described are similar to a laboratory protocol (the series of steps involved in carrying out a laboratory procedure or experiment) in that there is a specific analysis to be conducted at each step, and the results are added together to reach the final conclusion. Thus far, the procedures do not clearly match the traditional models of information behaviour.

FUTURE RESEARCH

The findings reported here are preliminary. More interviews must be conducted (for a total of twenty), to collect sufficient data to build a model of how bioinformatics resources are used to conduct a functional analysis of a gene. As well, the model itself must be tested and validated, both during its development, and once all of the interviews have been completed. The testing will be an iterative process, with the results of the initial testing being used to guide the next round of interviews.

Once the model has been developed and validated, we will proceed to test its effectiveness at facilitating access to bioinformatics resources for bench scientists. If found to be effective, then the model can be used as the basis for training and education programs, or as the foundation of a software tool

This research has studied how experts use bioinformatics resources. It would be interesting to also study those who are not expert in bioinformatics. It is possible that their patterns of use would be less consistent and focused, and involve more exploration and browsing of resources, and examination of the purpose of a search, following a more typical information seeking behaviour pattern.

The fact that clear patterns of use are already emerging from the data from a small number of interviews is promising, and is indicative of a unique information-seeking style among expert users of bioinformatics resources. It is important to continue this research to further understand and model the search patterns, and to explore how the model can be used to facilitate access to bioinformatics resources for bench scientists. Given the central role that bioinformatics will play in the future of biomedical research, it is essential to continue to explore and understand how these resources are used, and what can be done to make them more accessible.

ACKNOWLEDGEMENTS

The work reported in this paper is part of the author's dissertation research, under

the supervision of Dr. Elaine G. Toms and Dr. Joan M. Cherry, Faculty of Information Studies, University of Toronto, and Dr. A. Jamie Cuticchia, Bioinformatics Supercomputing Centre, Hospital for Sick Children. This research is partially supported by an NSERC grant held by E. Toms.

REFERENCES

Baltimore, David. 2001. Our genome unveiled. *Nature* 409:814-816.

Baxevanis, Andreas D. 2000. The Molecular Biology Database Collection: an online compilation of relevant database resources. *Nucleic Acids Research* 28 (1):1-7.

Bayer, A. E., and G. Jahoda. 1981. Effects of online bibliographic searching on scientists' information style. *On-Line Review* 5:323-33.

Collins, Francis S., Ari Patrinos, Elki Jordan, Aravinda Chakravarti, Raymond Gesteland, and LeRoy Walters. 1998. New goals for the U.S. Human Genome Project: 1998-2003. *Science* 282 (5389):682-9.

Curtis, K. L., A. C. Weller, and J. M. Hurd. 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., A. C. Weller, and J. M. Hurd. 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, Beverlee A. 1990. User needs and library services in agricultural sciences. *Library Trends* 38:415-41.

Garvey, W.D., K. Tomita, and P. Woolf. 1974. The dynamic scientific information user. *Information Storage and Retrieval* 10:115-31.

Grefsheim, Suzanne, Jon Franklin, and Diana Cunningham. 1991. Biotechnology awareness study, part 1: where scientists get their information. *Bulletin of the Medical Library Association* 79 (1):36-44.

Hackos, JoAnn T., and Janice C. Redish. 1998. *User and Task Analysis for Interface Design*. New York: Wiley.

MacLean, Marlie, and Colin Miles. 1999. Swift action needed to close the skills gap in bioinformatics. *Nature* 401:10.

Palmer, Judith. 1991a. Scientists and information: I. Using cluster analysis to identify information style. *Journal of Documentation* 47 (2):105-29.

Palmer, Judith. 1991b. Scientists and information: II. Personal factors in information behaviour. *Journal of Documentation* 47 (3):254-75.

Persidis, Aris. 1999. Bioinformatics. *Nature Biotechnology* 17:828-30.

Rolinson, J., A. J. Meadows, and H. Smith. 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, Meredith. 1999. Human Genome Project aims to finish "working draft" next year. *Nature* 398:177.

Yarfitz, Stuart, and Debra S. Ketchell. 2000. A library-based bioinformatics services program. *Bulletin of the Medical Library Association* 88 (1):36-48.