BIOLOGISTS' INFORMATION SEEKING BEHAVIOR WITH ONLINE BIOINFORMATICS RESOURCES FOR GENOME RESEARCH

By
Dihui Lu

A Master's Paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in Partial fulfillment of the requirements
for the degree of Master of Science in
Information Science

Chapel Hill, North Carolina

January, 2003

Approved by:

____

Advisor

Dihui Lu. Biologists' Information Seeking Behavior with Online Bioinformatics Resources For Genome Research. A Master's paper for the M.S. in I. S degree. January, 2003. 44 pages. Advisor: Gary J. Marchionini

A questionnaire survey of biologists was conducted to gather data regarding how biologists access and use online bioinformatics resources for their genome related research. The purpose of this study was to gain a better understanding of the information seeking behavior of biologists. A total of 57 respondents from academia, industry and government voluntarily participated in this survey. The survey indicated that the majority of biologists believe that online bioinformatics resources play very important roles for their research and show positive attitude toward future bioinformatics usage and training. These respondents are active users and confident about themselves in using online bioinformatics resources. Most have the basic skills to find information resources, and formulate queries. The results also revealed the information challenges posed by online bioinformatics resources such as how to keep up to date on information resources, hot to query over multiple resources and various training needs for bioinformatics applications.

Headings:

Bioinformatics

Biologist

Survey

Information seeking

Table of Contents

1. Introduction ……………………………………………………….………..…….... 1

2. Background and Literature review ……………………………..……..………... 2

2.1 Background ……………………………………………………………………... 2

2.11Bioinformatics……………………………………………………….…. 2

2.12 How to locate online bioinformatics resources ……………………………… 4

2.13 The heterogeneous database problem ……………………………….….... 5

2.14 User information interaction …………………………………………….... 6

2.2. Literature review …………………………………………………………… 7

3. Research design …………………………………………………………..…. 12

3.1 Technology choice ………………………………………..……….…..... 12

3.2 Selection of respondents ……………………………………………...….. 13

3.3 Distribution of the survey ……………………………………………..…. 13

3.4Questionnaire development ……………………………………..……..….. 14

4. Research findings ………………………………………………………..….. 15

4.1. Description of the user ……………………………………………………..... 15

4.2 User experience of using bioinformatics sources and services ………………... 19

4.3 Future needs for online bioinformatics resources 25

5. Summary and Conclusions ……………………………………..……….. 28

Bibliography ……………………………...…………………….…………… 32

Appendix I: Questionnaire …………………………………..……………..…. 34

Appendix II: Attachment for AA-IRB Proposal Form …………………………… 38

Appendix III: Invitation Letter ……………………………………….….……… 39

List of Tables

Table 1. working environment for the participants ………………………………. 15

Table 2. nature of work for the participants …………………………………….. 15

List of Figures

Figure 1. Self-reporting skills of biologists with different working environments 16

Figure 2. Self-reporting skills of biologists with different working nature …………… 17

Figure 3. Frequency of user access online bioinformatics resources 18

Figure 4. Self-reporting skills of biologists with different frequency of usage ……….. 18

Figure 5. How difficult to find the information sources ……………………………….. 19

Figure 6. The ways users find out information sources they need …………………….. 20

Figure 7. The way that biologists access online bioinformatics resources ………….. 21

Figure 8. Tools used for a task ……………………………………………………… 21

Figure 9. Keep up to date with bioinformatics resources …………………………….. 22

Figure 10. Type of query that biologists ask of online bioinformatics resources ……. 23

Figure 11. Ease of formulating queries in databases ………………………………… 23

Figure 12. Comparison of online bioinformatics resources as related to presentation, authority, reliability, speeds, query ability across multiple sources and interactivity … 24

Figure 13. The importance of bioinformatics resources for biologists ………………. 25

Figure 14. Prediction of future personal use of bioinformatics resources …………… 26

Figure 15. Willingness to attend bioinformatics workshop 26

1. Introduction

The volume of data produced in various genome projects has grown exponentially in last decade. The time has come for biologists to work on extracting knowledge from the huge amount of data. The information environment of biologists is changing rapidly due to new discoveries in genomics as well as developments in information technology. Online bioinformatics resources have dramatic impact on the way that biologists communicate and share resources for their research. The range of available technologies for biologists has expanded enormously and biologists are becoming intensive users of the electronic information resources such as computers, software, networks and databases.

There are two reasons for putting biological data on the Internet: retrieval and discovery. Retrieval is basically being able to get back what was put in. Amassing sequence information without providing a way to retrieve it makes the sequence information, in essence, useless. At the same time, what would be more valuable is to be able to get back from a system more knowledge than was put in by using information to make biological discoveries. Therefore, the biological data must be defined in a way that is amenable to both linkage and computation (Bioinformatics, A practical guide, 2000). A vision of bioinformatics is to help scientists make discoveries by discerning connections between two pieces of information that were not known when the data were entered separately into the same or different database or perform computations on the data that offer new insight into the records. The online bioinformatics resources have raised expectations of the scope of information that should be available electronically and how information is delivered. The capacities to search online biological databases, submit and download huge amounts of data, and perform analysis quickly and easily from government or organization sites are changing the expectations and abilities of biologists (Simmon, 1999). The biologists are gaining the skills to find more information, at a higher level of value to their work, in more customizable formats and at faster speeds (Simmon, 1999).

While the new information environment is really exciting and promising, it also provides the scientific community with serious challenges. Information seeking involves a number of personal and environmental factors especially in electronic environments (Marchionini, 1995). Suppose a biologist has a clear information need in mind, how can s/he determine: where to begin? What resources are helpful? Where the information resources are available? How to use those information resources? These are all interesting questions that need to be addressed. In order to answer these questions, we need to know more about the information seeking process of biologists. This paper presents a descriptive study of online bioinformatics resources access and usage as part of the broader information seeking activity of biologists for their genome related research.

2. Background and Literature review

2.1 Background

2.11 Bioinformatics

There are different ways to define bioinformatics. On the NCBI (National Center of Biotechnology Information) website, Bioinformatics is defined as “a new discipline which is merged from biology, computer science, and information technology”. “Bioinformatics researchers try to develop and apply computing tools to extract the secrets of the life and death of organisms from the genetic blueprints and molecular structure stored in digital collections” (NCBI website). From the information science perspective, bioinformatics may be defined as the acquisition, analysis, utilization, storage and retrieval of massive amounts of biological sequences, structural data and the associated annotations etc. Therefore, it is a field that marries information management techniques with an understanding and appreciation of the significance of biological data (Sobral B. 2000). It is believed that bioinformatics will deal with the new challenges in biology and allow the new century of biology to bear fruit.

Traditionally, biology research begins with a hypothesis. A biologist then collects experimental data and analyzes them to support or disprove the hypothesis. However, bioinformatics is changing this sequence of events which is leading to a change from experimental science to discovery science (Sobral B. 2000). Today, large-scale exploratory experiments allow scientists to gather as much data as possible automatically. The Human Genome Project, for example, is creating an inventory of all 3 billion amino acids in the human genetic blueprint. Besides the exponential growth of data, new types of data emerge regularly, data are updated very frequently, accessed intensively and exchanged very often by researchers on the Internet (Frédéric A. et al., 2000). It might be possible that when a biologist forms a hypothesis, the result may already be in such a data collection, just a computer search away. Therefore how to help biologists to store, retrieve and annotate these data effectively are of great importance for bioinformatics study. In general, there are three interesting questions related to how biologists access and use online bioinformatics resources. These questions are considered in turn below:

2.12 How to locate online bioinformatics resources?

The competitive and developing nature of biological research and biotechnology industries are highly dependent on up-to-date information. Over the last two decades, the development of high throughput data generation factories and novel laboratory technologies such as large scale sequencing, proteomics and microarray analysis, has transformed biology from data poor to "data poisoned" quickly (Sobral B. 2000). Consequently, the bottleneck for biological research has shifted from data generation to data management. Because of such challenges and bottlenecks, biology in the 21century is being transformed from a purely lab-based science to an information science as well (NCBI). As more and more genomes including the human genome draft have been determined. Finding ways to take advantage of the huge amount of information generated is really challenging. The operative principle most prominently involved in transmitting the fruits of genomics has been open access (Varmus H. 2002). The availability of the sequences of many genomes through the Internet is making an extraordinary amount of essential information freely accessible to anyone with a desktop computer and a link to the World Wide Web. But the information itself is not enough to allow efficient use. It is important for people to know where best to find the information resources and the software to perform retrieval and analysis.

Nowadays, a typical bioinformatics retrieval system has to deal with information volumes up to one terabyte and information resources that are in a distributed and heterogeneous format on the Internet. The Molecular Biology Database Collection, for example, currently holds over 500 information resources including 281 key databases (Baxevanis, 2001). In addition, a biological database is not just a big collection of data, there are many new computer software and tools associated with those information resources in order to help people to retrieval and annotate the data. Therefore, it is a challenging job to provide researchers with fast and efficient access to data and information and to provide computational assistance to molecular biologists in analyzing their local data. Given the quantity of the online information resources and importance of them to biological research, find how biologists locate the bioinformatics resources and tools needed for their research and how they stay well informed of information resources are all very interesting questions for bioinformatics researchers.

2.13 The heterogeneous database problem.

Another active study area of bioinformatics is the heterogeneous database problem because all kinds of data are distributed in hundreds of heterogeneous databases in different formats. Because of such a problem, biological databases have mostly served as a "memory" function for the biological research community (Sobral B., 1999). However, simply storing data in a database does not provide biological researchers with the needed context for those data to be truly useful in the discovery of new biological knowledge, rules, or principles (Sobral et al., 1999). Multiple types of data must be queried and compared together for biologists to discern the inherent relationships between the data. Numerous labs and organizations built their own database, data warehouse for their own purpose during the last two decades. Most of the information systems were built in an *ad hoc* way without systemic thinking that may create problems later for data integration. These information resources were developed over long periods of time using various proprietary technologies. There are duplicated, conflicting or even erroneous information in different data repositories which results in unnecessary constraints for utilization of the information and its transformation into knowledge and products.

Generally speaking, in biological information management, people suffer from two levels of heterogeneity: One is querying across different systems housing the same types of information, for example, genetic maps in RiceGenes and MaizeDB; and a second is querying across different types of data that need to be related and available for analysis through a single interface (Sobral, 2001). For example, plant breeders need information for genetic maps, DNA sequences, gene expression profiles and phylogeny information for crop improvement. However, the required information is distributed in separated and heterogeneous databases and there is no single unified interface to provide all the information. Therefore, the problem of using information housed in different databases providing different types of access is under heavy study by bioinformatics researchers.

2.14 User information interaction

While the problems of managing biological information have not been solved satisfactorily, how users express and issue their queries against the online bioinformatics resources, how online bioinformatics resources help user to perform their queries are also problems that need to be addressed. As mentioned above, hundreds of biological databases and online tools have been established to facilitate genome research. All kinds of users including biologists, programmers and database managers need to issue complex queries through the Internet or locally by command line. Therefore, the query ability and query presentation will have deep impact on the degree to which the information resources are utilized. All these considerations are relevant for user interface design and data mining via information visualization. Therefore, as a starting point, it is very important to understand how user and online information resources interact with each other. That is where information science comes into play. A better understanding of user requirements is an important part of the bioinformatics application design process especially for user interface design and database development (Stevens R., et al., 2000). This could be achieved partially by the study of the human information interaction that takes place as biologists search for bioinformatics information. Then, the user needs generated could be incorporated into the software design process and used for usability studies to check the usefulness and effectiveness of the tools (Stevens R., et al., 2000).

2.2. Literature review

There has been a notable lack of research as it pertains to the information needs and information seeking behavior of the biologists. The publications for user study directly related to bioinformatics are quite limited. Given the wide variety and diversity of information resources available to biologists as well as the importance of such knowledge in their decision making during their research, an investigation of how biologists access and use online bioinformatics resources is necessary.

In general, user study is a very important area under most research in library and information science and a large body of literature can be found in this discipline. Wilson suggests that the information seeking behavior results from recognition of some need perceived by the user (Wilson, 1981). A user may try to locate and acquire the information desired following formal and/or informal channels of communication and the process that a user will engage in determining the information seeking behavior (Siatri, 1998). The origin and conception of user studies were based upon the belief "that if one could somehow identify the information needs and uses of a population subset, one could design effective information systems"(Crawford, 1978). Therefore, studying a community and its needs helps to provide better information services on demand.

Methodologies for user studies have been under debate for a long time among information scientists. There are two dominant types of user studies, namely, the system-oriented studies and the user-oriented studies. The system-centered approach views the users as passive information recipients and investigates their external behavior, generally by means of quantitative methods (Siatri, 1998). Although these studies yield an overall picture of information needs and seeking behavior, they fail to convey a real picture reflecting the factors which trigger the information search and a more in-depth insight into the individual's conception's and thoughts (Siatri, 1998). On the other hand, user-oriented studies view the users as active and self-controlling recipients of information. These studies are concerned with the internal cognitions of users and are investigated by qualitative methods (Dervin and Nilan, 1986). User-oriented methods take users needs into account in order to create an environment which will be friendly, effective and easy for users to use. Many researchers support user-oriented study because the conceptual framework upon which user-oriented studies are based acknowledges the dynamic and responsive nature of human behavior and thus of information seeking (Siatri, 1998). Despite the large number of user studies that have been conducted, our knowledge as far as it concerns user needs and information seeking behavior is far from enough. Due to the rapid changes in the electronic information environment, the lack of understanding the information needs and information seeking behavior poses an obstacle for the delivery of electronic information service. Thus, user studies have been started for many other disciplines such as the biology community to acquire a more in-depth understanding of the users of information needs and behaviors.

There are several studies have been done to address the problem of online information resources usage and information needs in cancer research. For example, Bult C. J. did a descriptive survey on web resources for basic cancer genetics research and summarized the information resources and databases available on the World Wide Web for cancer research (Bult C. J., 1999). In another study, Lomax et al worked with people from library science and medical informatics and identified some factors that may motivate oncologists to seek information and accurately describe the information seeking behavior of these oncologists (Lomax E. C., et al. 2000). The problems they tried to address included what are the information needs of medical oncologists? What are the current problems medical oncologists facing in meeting their information needs? Are these needs being met through the use of information resources such as textbook, online resources and databases, and journals? How can evolving information resources and technologies help the oncologists meet his or her information needs? They used a multimethod approach of mail survey, structured observation and personal interview of practicing oncologists to study the questions above. Some interesting findings have been identified such as: an oncologists comes to an information source as part of a way to build or rebuild his or her knowledge; and oncologists want information that is high quality, available, accessible, concise and organized to support clinical decision making. This investigation of the information needs of medical oncologists and their information behavior provided a clearer picture of the information world of the oncologists which may aid clinical decision support, continuing education and patient care in the future.

There have been a number of studies on how British biologists use computers for information handling. A.J. Meadows et al. did a comparative study on how scientists use information technology for their research in British and Saudi Arabian universities (Bukhari A.A. and Meadows A. J., 1992). Most of these researches are focusing on surveys of computer usage. In 1995, A. J. Meadows did an in-depth study on the use of information technology by biological researchers (Smith, H., 1995). It surveys the usage of information technology and related factors by biological researchers at four institutions. They found some interesting results such as there are difference in usage depending on the institution and fields involved; the senior researchers in biology are typically more information active as information providers and recipients than junior researcher which could be explained in terms of the pressures on senior staff time and fewer financial restrictions. Overall, their research indicated that, though the information-handling activities of biologists may differ on average from other fields, the differences in computer-based information-handling within biology are as wide as anything to be found across the sciences as a whole. In another study, a paper-based survey was conducted to classify the bioinformatics tasks currently undertaken by working biologists (Stevens R., et al., 2001). The questionnaire survey was distributed to biologists both in academia and industry to gain a representative set of queries and tasks that need to be supported and the components needed to implement in a general query system. Some sample questions include: what tasks do you most commonly perform? What tasks do you commonly perform, that should be easy, but you feel too difficult? What questions do you commonly ask of information resources and analytical tools. The study is a good starting point to study user requirements, however, this study got only 35 respondents and the survey was paper based.

As discussed above, to date, little formal work has been done to investigate the significance of new information technology such as online bioinformatics resources on the information needs and information seeking behavior for biologists. Therefore, more effort should be made to investigate how biologists access and use online bioinformatics sources and services, those facets such as how do they find out about the information resources; how do they formulate their queries against those online resources; what queries users want to be able to ask the information sources and what visual ability users want when the information systems present the query results. This preliminary survey aimed to inform future studies on interactive interface design, data mining and biological databases development and interoperation.

3. **Research design**

A questionnaire survey of biologists, in academia, industry and government was taken to create a picture of the usage and attitudes of biologists toward online bioinformatics resources. The survey was carried out via the Internet from February to March in 2002. It was organized in three parts: user profile, user experience with online bioinformatics services and future needs. Scales and index are used to compose the survey questions. The purpose of the study was to gather descriptive data regarding:

- Pattern of information-seeking by biologists who are working on genome related research in regard to online bioinformatics resources. For example, user profile such as frequency and attitudes toward online bioinformatics resources that may influence use pattern.
- Problems that biologists are facing regarding online bioinformatics resources.
- Variables that may affect these information behaviors such as user background, self-reported information seeking skills, personal attitude toward usage of online information resources and future training needs.

3.1 Technology choice

The web-survey was anonymous and all the information the user presents was saved to a database. Coldfusion was used to generate the online survey and trace where the user is from based on their IP address if the participant was unwilling to provide their physical address. The purpose of keeping track of the IP address of the respondents was only to identify the unique reply.

An Access database was established to store all the survey responses provide and the user environment variables (IP address, date…).

3.2 Selection of respondents

The survey was sent to the biologists selected from the members of The Arabidopsi*s* Information Resources (TAIR). TAIR is one of the major genome research communities. The *Arabidopsis* is one of those model organisms that have been sequenced such as *Yeast, E. Coli, Drosophila*. TAIR is the official site which provides a comprehensive resource for the scientific community working with *the Arabidopsis thaliana* genome. In order to get better representative results from the survey, the participants was chosen from biologists whose predominant nature of work are related to genome research including genome sequencing, sequence annotation, pattern analysis and function analysis. They included working faculties, graduates or working professionals either in academics, industries or federal government. The survey was distributed to the members of TAIR via a web form.

3.3 Distribution of the survey

The questionnaire was distributed to biologists via email. A collection of email lists was generated by querying TAIR membership web database. Since the general return rate for such a survey is pretty low, 450 email invitations were sent out to get as many respondents as possible. A Perl script was written to send bulk emails with the customized content for each invitation.

3.4 Questionnaire development

The self-reporting questionnaire methodology was used because it has several advantages. Self-reporting is a good way to measure individual attitudes (Robson, 1993). This questionnaire has a web interface which contained the survey questionnaires, invitations, references and collection of online bioinformatics resources for appreciation. Questions were developed based on the one used by Steven R. et al for their study on classification of bioinformatics tasks. (Steven R. et al., 2000). Closed questions are used because they are easier for participants and simpler to analyze than open-end ones.

4. Research findings

4.1. Description of the user

There were 450 invitations sent out, however, there were about 100 email invitations returned due to expired email addresses. Fifty-seven valid questionnaires were returned and the overall response rate is 16%. The number of respondents is shown in the following table 1:

Table 1: working environment for the participants

Question: Which of the following best describe your working environment?	Academic	40	70%
	Industry	13	23%
	Government	4	7%
Total		57	100%

Among the survey participants, 70% of people were academe, 23% of the people were from industry and 7% of respondents were from federal government research organizations. Since the majority of the members in the TAIR are from universities, research institutes, the members from industry rank the second and with a few members from federal government research agencies, the distribution of the participants correspondents with the general distribution of members from different working environments.

Table 2: nature of work for the participants

Question: Which of the following best describes the nature of your work?	Large scale sequencing	0
	Functional analysis of genome	38
	Genome bioinformatics	9
	others	14

As shown in table 2, the participants work on different projects. Most of people (about 66%) are working on functional analysis and annotation of the genome, about 14% of people are working on genome bioinformatics. Others are working on general molecular biology, functional analysis of novel proteins, molecular genetics, metabolic engineering, molecular phylogenetics, cell biology, and gene transformation. No respondent reported working on large-scale sequencing was reported. One reason could be, perhaps, with more model organisms and other plant genomes have been sequenced, more and more researchers have been working on annotation or functional analysis of the date generated from the genome projects.

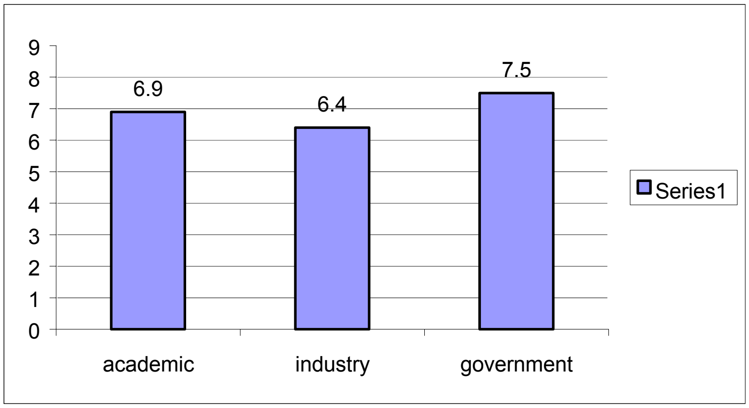


Figure 1: Self-reporting skills of biologists with different working environments
(1: never run a blast search 9: confident with running queries across multiple sources)

As shown in figure 1, all participants have prior experience with online bioinformatics resources. The self-reporting skill levels ranged from 1 to 9 on a nine-point scale. Overall, the skill levels were high. The skill levels for respondents with different working environment from academic, industry or government are very close with average scores 6.9, 6.4, 7.5 respectively. Therefore, we can see that the bioinformatics resources generally deal with a relatively homogeneous community with similar skill level. The possible explanation is that all biologists are active users for online bioinformatics resources and they are pretty familiar with the information sources they frequently use. Another reason is probably because the working environments for biologists are quite homogenous regardless of wherever they work at universities, companies or government organizations.

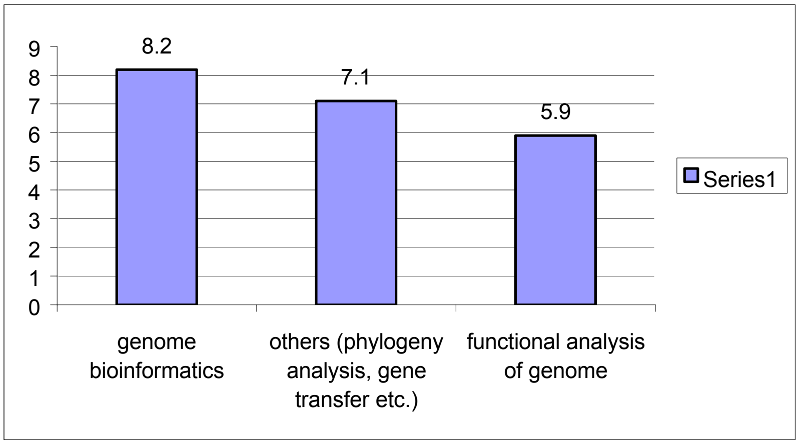


Figure.2: Self-reporting skills of biologists with different working nature
(1: never run a blast search 9: confident with running queries across multiple sources)

However, the skill levels varied for people with different nature of work as shown in figure 2. People work on projects related to functional analysis of genome scored only 5.9 while people work on genome bioinformatics scored 8.2. Biologists who work on phylogeny analysis, cell biology or other fields are in the middle with an average score 7.1. This result indicated a skill gap between general biologists who focus on using bioinformatics tools and genome bioinformatics researchers who more focus on development of new software and algorithms. Therefore, how to give general biologists more training to improve their skills is very important in the future.

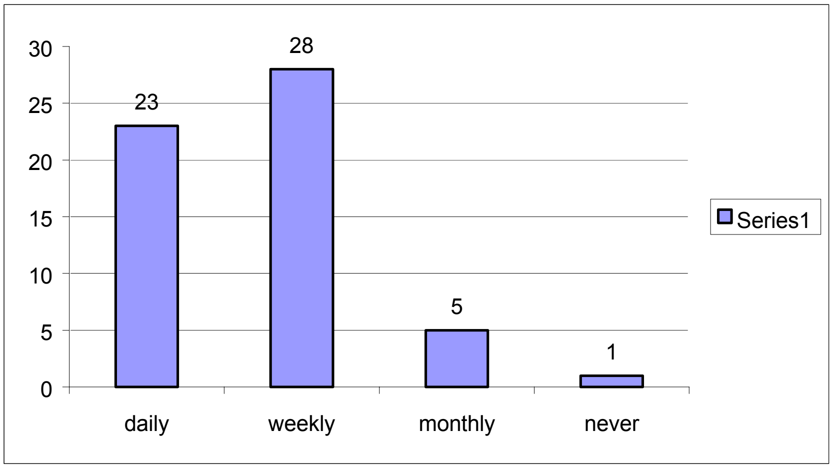


Figure 3: Frequency of user access online bioinformatics resources

In order to get a better understanding of how often the biologists use online bioinformatics resources, the participants were asked to report their frequency of visiting and using those resources. Overall, figure 3 shows that 90% of the respondents have been using online bioinformatics resources daily or weekly, and only 9% of people access and use online bioinformatics resources monthly. This result shows that bioinformatics resources have become an inseparable part of biologists' routine research.

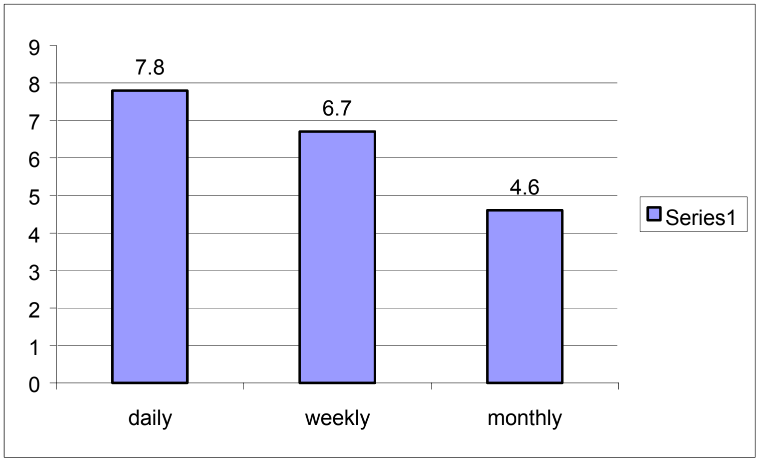


Figure 4: Self-reporting skills of biologists with different frequency of usage
(1: never run a blast search 9: confident with running queries across multiple sources)

As shown in figure 4, respondents use online bioinformatics resources on a daily basis have higher skill level, and the skill level for the weekly users is in the middle while monthly users score lowest.

4.2 User experience of using bioinformatics sources and services:

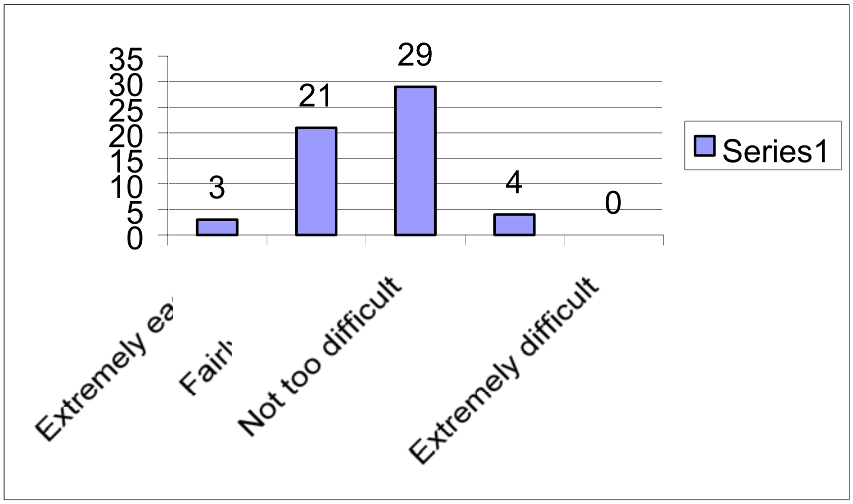


Figure 5: How difficult to find the information sources

Figure 5 shows that about 50% of biologists feel that finding online bioinformatics resources in general, is not difficult. 37 percent reported fairly easy and 3 percent reported it is extremely easy. Only 4 percent of biologists have difficulty finding the online bioinformatics resources they need. This result is somewhat surprising because it seems most people know what they need and where to find it. One possible explanation could be that this community is quite knowledgeable and highly educated. Another possible reason might be the general problem with self-reporting experience in which users may over-estimate themselves. While the biologists are quite familiar with their own research area and information resources, they may be dependent on several major resources without knowing that there might be better sources for their purpose.

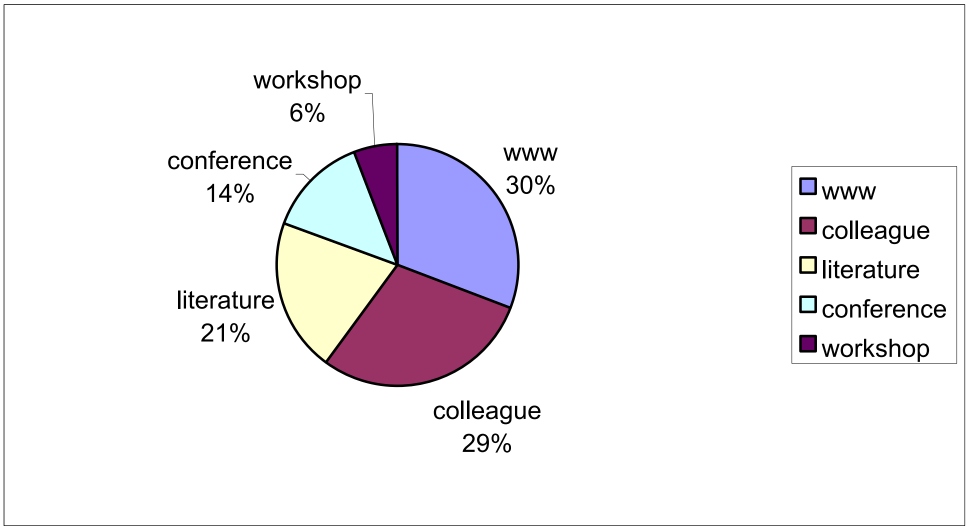


Figure 6: The ways users find out information sources they need

Figure 6 shows that respondents learned about online bioinformatics resources from the World Wide Web (30%), colleagues (29%), literatures (21%), conferences (14%) and workshop (6%). Most people use the World Wide Web to find out the information s/he needs because of its speed of access to information and the scope of information available. Further it indicates that the World Wide Web has played so important role that had changed the way that people do research, communicate to each other. Meanwhile, from the survey, we found that about 29 percent of people are dependent on their colleagues at work to find information resources. This result corresponds with the finding that people always try to get information from people in similar situations as themselves.

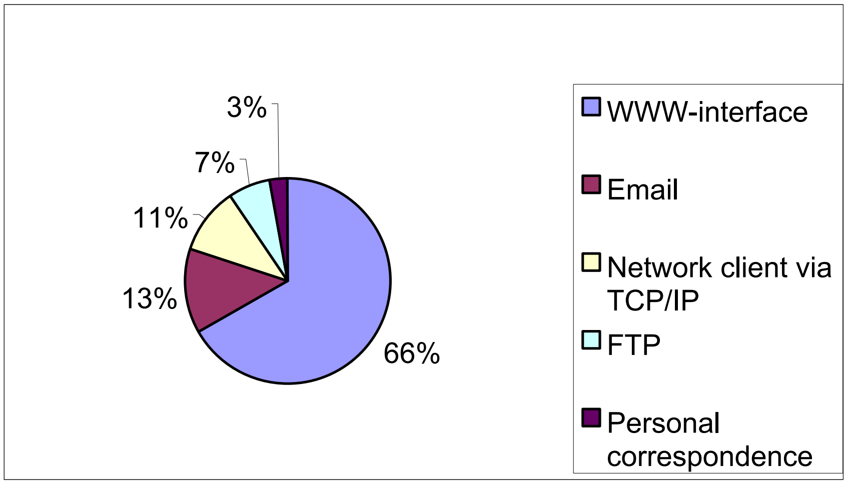


Figure 7: The way that biologists access online bioinformatics resources

It is also important to figure out the ways by which biologists access the online bioinformatics resources. Based on figure 7, 66% of the respondent access online bioinformatics resources through a WWW-interface, 13% use email, 11% use a network client via TCP/IP, 7% use a FTP client and only 3% of people access information resources with personal correspondence. The result suggested that the World Wide Web-interface has become the major way that biologists access online bioinformatics resources. Therefore, good interface design is an important factor that may affect the efficient use of those resources.

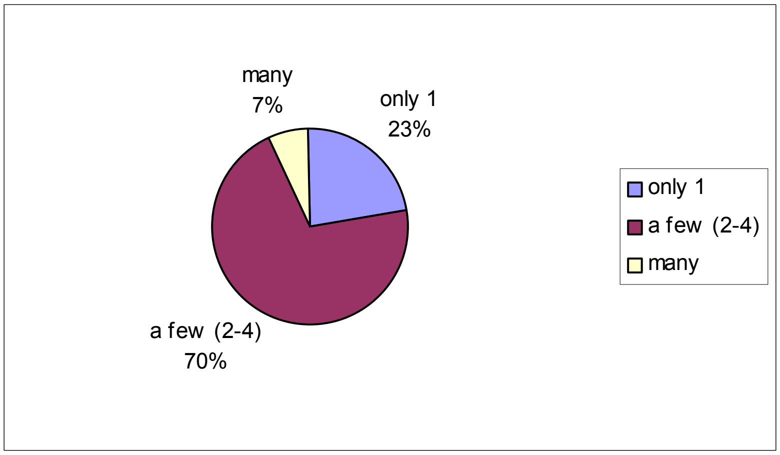


Figure 8: Tools used for a task

Compared to databases in other disciplines, biological databases are very specific. It is not just a big collection of data but also associated with some bioinformatics software and tools for people to analysis the data. Therefore, it is important to know how many tools a user generally uses for a task such as a functional motif search for a protein or homology alignment. Among all the respondents, about 70% of respondents reported using 2-4 tools, 23% use 1 tool and only about 7% use many tools (more than 4) as shown in figure 8.

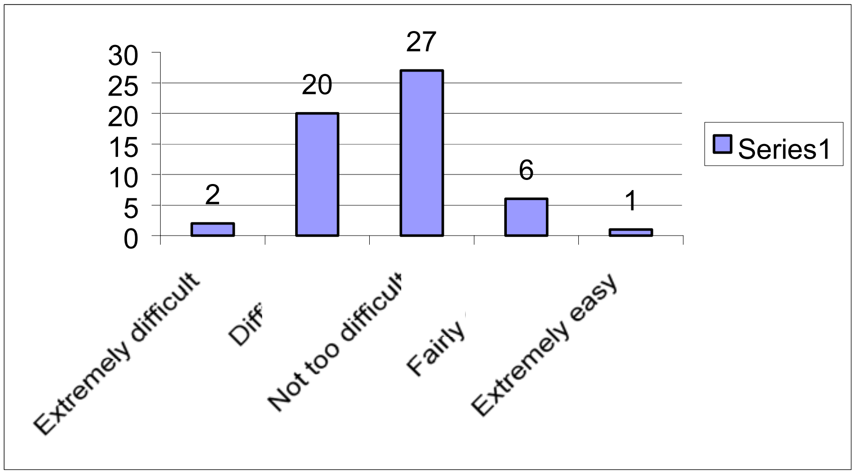


Figure 9: Keep up to date with bioinformatics resources

As we already know that biologists are dependent on up to date information to make their decisions, therefore, how biologists keep themselves well informed of the newest information sources is an interesting question that need to be addressed. Based on the survey results, 47% of biologists feel it is not too difficult, 35% feel difficult, and two people feel extremely difficult to keep up to date with the current bioinformatics resource. Only 10% of the respondents think it is fairly easy for them to keep up with current information resources (see figure 9). This result indicates that the majority of users still have difficulty to keep up to date with the most recent current bioinformatics resources since they are updated so frequently.

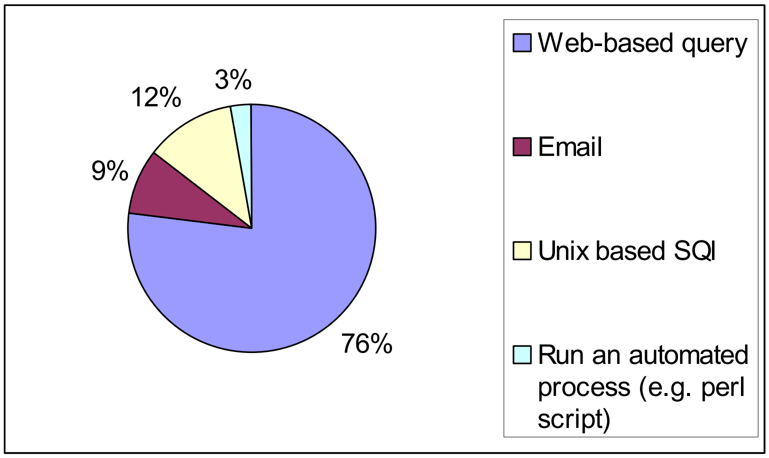


Figure 10: Type of query that biologists ask of online bioinformatics resources

As shown in Figure 10, basically, there are four types of queries that biologists generally ask an online bioinformatics resource, web-based query (76%), email (9%), unix based SQL (12%) and run an automated process such as Perl script (3%).

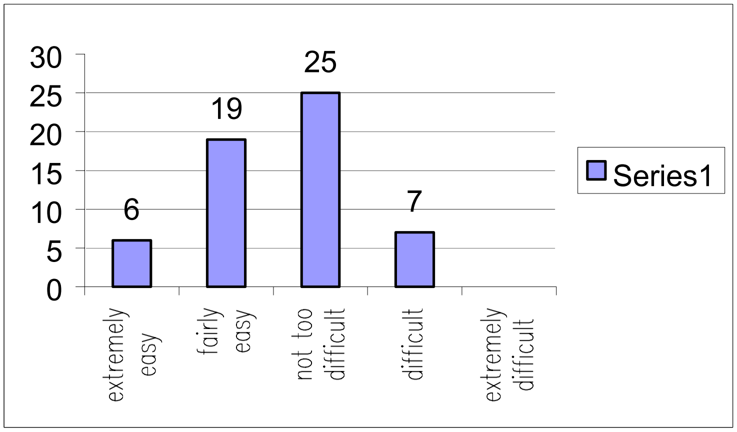


Figure 11: Ease of formulating queries in databases

Since query formulation involves matching understanding of the task with the system selected (Marchionini, 1997), it is necessary to understand how users formulate their queries in the first place. Based on survey results shown in figure 11, 43% of biologists feel it is not too difficult to formulate their queries, 33% feel it is fairly easy and 10% said it is extremely easy. Only 7 persons felt it is difficult to formulate their queries against the online bioinformatics resources. This result indicates that most of biologists have a clear target in mind when they want to ask for bioinformatics resources.

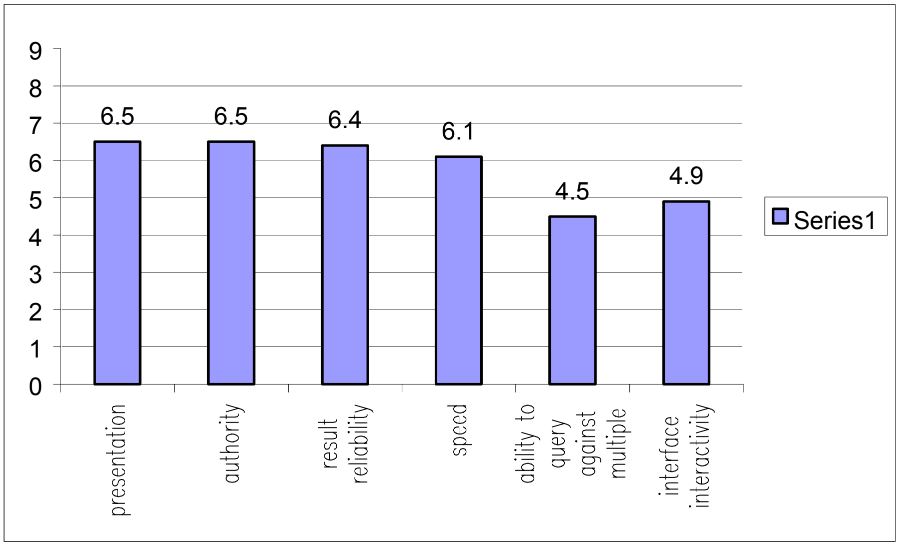


Figure 12: Comparison of online bioinformatics resources as related to presentation, authority, reliability, speeds, query ability across multiple sources and interactivity

In order to get an overall picture about how users are satisfied with the online biological databases, the participants were asked to score an interface of their choice in terms of its general presentation, information authority, result reliability, speed to process their queries, ability to query multiple resources and its interactivity. From the survey results shown in figure 12, the presentation, authority, result reliability and speed scored 6.5, 6.5. 6.4, 6.1 respectively which are much higher than ability to query across multiple resources (4.5) and interactivity (4.9). This result indicated that the users are generally satisfied with online biological database interface presentation and speed, and they trust the information authority and reliability. The major problems with those databases are to provide the ability to allow users to query against multiple resources and to allow users to change experiments and parameters to change the results. This result is correspondent with the additional comments made by some survey participants. For example, one respondent said he would like to see improvements in Map Viewer to allow seamless access to DNA sequences from the map. This raises an interface question about how to allow users to query both a genetic map database and a sequence database without prior knowledge. Another participant said the parameters for the user to set up are too complicated in bioinformatics tools, and it is difficult for a biologist to understand and set them up properly. Since there are so many bioinformatics tools available, different algorithms and statistical models are used behind those tools, thus it would be extremely difficult for a biologist to change the parameters to do data mining without training in computational biology. It would be a dangerous thing if biologists just use those bioinformatics tools without understanding the underlying algorithms and parameters. Therefore, the training in use of bioinformatics tools will help biologists take more advantage from those information resources.

4.3 Future needs for online bioinformatics resources

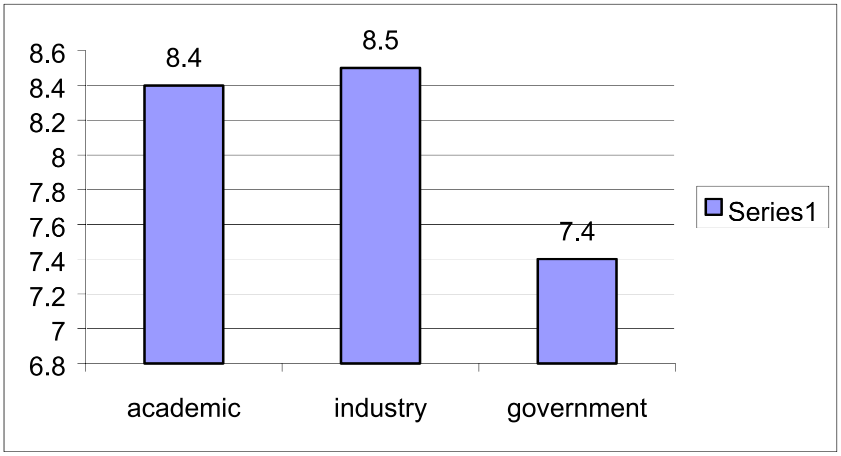


Figure 13: The importance of bioinformatics resources for biologists

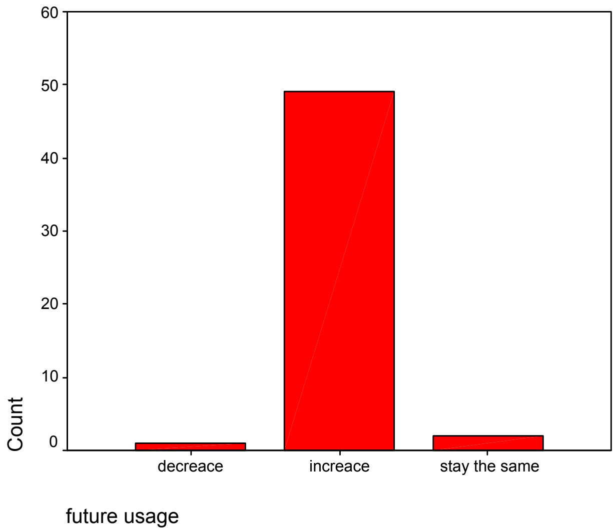


Figure 14: Predictn of future personal use of bioinformatics resources

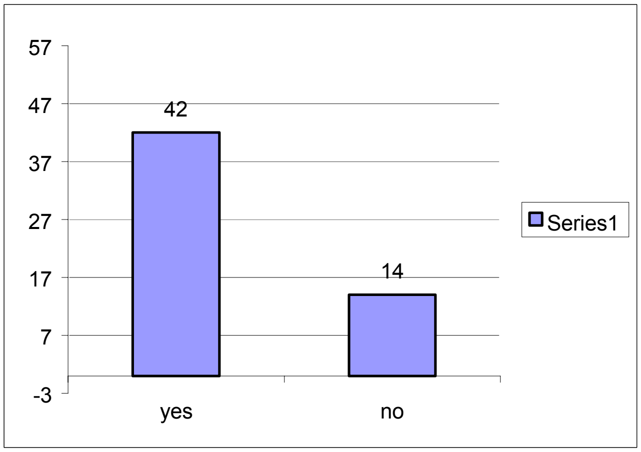


Figure 15: Willingness to attend bioinformatics workshop

In order to predict future bioinformatics usage from the biologists' point of view, the participants were asked three questions: How important is online bioinformatics resources in the advancement of your biological knowledge? How do you anticipate your personal use of online bioinformatics resources? Are you interested in attending workshops or taking classes in Bioinformatics? From figure 13, we can see, bioinformatics plays an important role in advancement of knowledge for biologists from academic area, industries and government agencies. Generally, the majority of participants believe their usage for online bioinformatics resources will increase (see figure 14). 75% of the people are willing to attend workshops or taking bioinformatics classes (see figure 15). These results indicate that more and more people have realized the importance of bioinformatics resources and it have become and will continue be an inseparatable part of biological research.

5. Summary and Conclusions

Biology is complex and so are the online bioinformatics resources. Due to it, we still do not understand their complexities and interactions. The molecular biology community is a distributed one with a culture of sharing substantial quantities of rapidly evolving information (Davison S. B., 2001). However, the development of a global informatics infrastructure to support this community has been piecemeal (Davison S. B., 2001). In order to develop better bioinformatics applications to support biological research, we need to gain a deeper understanding of the users, the biologists. From this survey, a big picture of biologists' use of online bioinformatics resources can be described as following: a substantial number of biologists work for academic institutions, industries or government research institutes. The majority of them believe that bioinformatics resources play very important roles for their research. Basically, they access and use online bioinformatics resources daily or weekly for genome related projects such as functional analysis of genome, genome bioinformatics and phylogeny analysis. These people are quite knowledgeable and confident about themselves in using online bioinformatics resources. They feel quite comfortable to formulate queries and find out where the information resources are. However, due to the rapid changing information environment, many people still have difficulty in keeping up to date with the information resources. Most people learn how to access and use information resources through the World Wide Web or from experienced colleagues at work. Most of the time, they query the online bioinformatics resources by web-based queries and email while a few people use some advanced type of query such as Perl scripts, SQL (structured query language). Generally speaking, the biologists are satisfied with presentation of the online database interface and speed, they trust the information authority and result reliability, but they would like to see more improvement on the ability of querying across multiple sources and improved interactivity of the query interface. Most of the biologists are willing to attend certain kinds of training such as classes or workshop on bioinformatics tools.

One major problem reflected from this survey is, as mentioned earlier, query across multiple databases. There is a strong indication from users that the inability to interoperate between tools was a barrier to asking more complex questions since each area of molecular biology generates its own databases, and a wide range of specialized interrogation and analysis tools are commonly used over these resources. It seems that performing searches and finding data are not difficult for biologists, the intelligent use of all of accumulated facts from databases is. Many biological data resources are frequently not databases in the conventional sense in that little distinction is made between databases (e.g., Entrez) and tools (e.g., BLAST). Many databases do not have a separate schema containing their meta-data or if they do, it is not freely accessible (Globe, C. A., 2001). Most are tools, processes (e.g., sequence alignment), or proprietary flat file structures containing embedded meta-data, with a limited set of parameterizable services accessed through a call-based interface (Globe, C. A., 2001). These resources are poorly integrated and difficult to use together. The characteristics of bioinformatics resources above have become significant drawbacks if we consider the complex retrieval tasks that biologists working in this environment are typically required to undertake. If biologists wish to go beyond the standard provision offered by predefined query systems such as NCBI Entrez, they must develop their own analysis program which is time and cost consuming. Therefore, how to build a unique interface and network that can combine dispersed researchers, computer resources and information into a single integrated computer and communication environment to provide the users the seamless access to multiples information resources, is a really challenging task. A number of approaches have been used so far, from Web-based browsers to data warehouses to integrate heterogeneous databases. These approaches includes: meta-data based approaches to provide transparent access to multiple resources, centralized approaches such as Gene Bank and federated approaches such as NCBI. In the near future, they will play a major role in helping researchers with their increasing access to databases residing on remote machines for the retrieval, analysis and sharing of data.

Another feedback from the survey is how to provide more training to biologists and help them benefit more from current bioinformatics resources. Training for understanding the algorithms behind the applications, making biological sense of the parameters set up, awareness of better information resources, are all of great importance. “The more you learn, the less you feel you know”. In the past, skilled colleagues and online training tutorials have contributed most to the use of bioinformatics resources for biologists. In the future, as a great supplemental factor, the training provided by bioinformatics professionals should and will play a more important role.

From the information science perspective, we believe that understanding user requirements is an essential step in designing future bioinformatics applications, such as databases, tools and user interfaces. Especially, when query based systems are designed, it is essential to know what range of queries to offer and the mechanisms needed for their support. This web survey was carried out to investigate how biologists use online resources for their genome research. We hope the feedback from the users have shed some light on the information seeking behavior of the biologists and the range of tasks that needs to be supported in a general query system of online bioinformatics applications. We hope that the survey results will be incorporated in future application design and evaluation, and therefore benefit both working biologists in the genome research community and the bioinformatics researchers in the long term.

Bibliography:

Andreas D. Baxevanis (2001)The Molecular Biology Database Collection: an updated compilation of biological database resources Nucl. Acids. Res. 2 29: 1-10.

Baxevanis A. D. et al., (2001) Bioinformatics:A practical guide to the analysis of genes and proteins. Second edition. Publisher: Wiley-Interscience.

Bruno W. S. Sobral (2000) Bioinforamtics and the future role of computing in biology. Available at http://www.agbiotechnet.com/proceedings/10_Sobral.pdf. Last viewed December 2nd, 2002.

Bukhari A.A. and Meadows A. J., (1992) The use of information technology by scientists in British and Saudi Arabian universities: A comparative study, Journal of information science 18 (5): 409-415.

Bult C. J., Krupke D. M., Tennent B.J., and Eppig J. T. (1999) A survey of web resources for basic cancer genetics research, Genome research. Vol 9 (5): 397-408.

Crawford, Susan. (1978) Information needs and uses. In: Williams, Martha, ed. Annual Review of Information Science and Technology: volume 13. Knowledge Industry Publications, p.61-81.

Davidson, S., Overton, C. and Buneman, P. (1995) Challenges in integrating biological data sources. J. Comput., Biol., 2, 557-572.

Davison, S., et al. (2001) K2/Kleisli and GUS: Experiments in integrated access to genomic data sources. Available at http://www.research.ibm.com/journal/sj/402/davidson.html. Last viewed December 2nd, 2002.

Dervin, Brenda and Nilan, Michael.(1986) Information needs and uses. In: Williams, Martha, ed. Annual Review of Information Science and Technology: volume 21. Knowledge Industry Publications, p.3-33.

Frédéric Achard, Guy Vaysseix, and Emmanuel Barillot. (2000) XML, bioinformatics and data integration Bioinformatics. 17: 115-125.

Globe C. A., Stevens R., Ng G., Bechhofer S., Paton N. W., Baker P. G., Peim M., and Brass A. (2001) Transparent access to multiple bioinformatics information sources.

Available at http://www.research.ibm.com/journal/sj/402/goble.html. Last viewed December 2nd, 2002.

Kraener E. and Ferrin T. (1998) Molecules to maps: tool for visualization and interaction in support of computational biology Bioinformatics. Vol (14) 9: 764-771.

Lomax E. C., Lowe H. J., Logan T. F., An investigation of the information seeking behavior of medical oncologists in metropolitan Pittsburgh using a multimethod approach. Available at http://www.amia.org/pubs/symposia/D004291.PDF. Last viewed December 2nd, 2002.

Marchionini G. (1995) Information seeking in electronic environments. Cambridge.

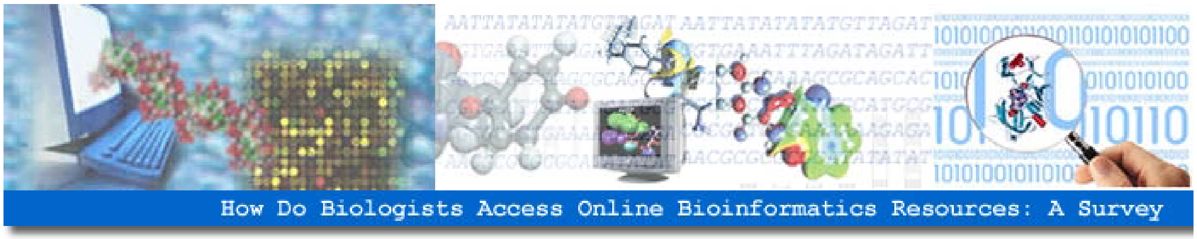
Siatri R.Information seeking in electronic environment: a comparative investigation among computer scientists in British and Greek universities. Available at Http://informtionr,net.ir/4-2/isic/siatri.html. Last viewed December 2nd, 2002.

Stevens R., Globe C. Baker P. and Brass A. (2001) A classification of tasks in bioinformatics. Bioinformatics Vol 17 (2): 180-188.

Sugar, Williams. (1995) User-centered perspectives of information retrieval research and analysis methods. In: Williams, Martha, ed. Annual Review of Information Science and Technology: volume 30. American Society of Information Science.77-109.

Wilson,T.D. (1981) On user studies and information needs. Journal of Documentation, 37(1): 3-15.

Appendix I: Questionnaire


How Do Biologists Access Online Bioinformatics Resources: A Survey

Project Description | Consent Information

. Section I---User Profile

1.Which of the following best describes your working environment:

□ Academic □ Industry □ Government Other

(Please specify:) □

2.Which of the following best describes the nature of your work:

□ large scale sequencing	□ functional analysis of genome
□ genome bioinformatics	Others (Please specify:) □

3. How would you assess your skill at using online bioinformatics resources such as **NCBI blast search, literature search, protein motif search, multiple sequence alignment, sequence assembly and contig analysis** etc.? Please choose the level that most closely matches your skill level. For example: Level 1 - never run a Blast search Level 9 - confident running queries across multiple resources

1 □ 2 □ 3 □ 4 □ 5 □ 6 □ 7 □ 8 □ 9

4. How often do you use online bioinformatics sources and services?

□ Daily □ Weekly □ Monthly □ Monthly □ Never

. Section II---Your experience of using online bioinformatics sources and services

5. How difficult has it been to find the online bioinformatics resources you need for your research?

□ Extremely difficult □ Difficult □ Not too difficult

□ Fairly easy □ Extremely easy

6. How do you find out the online bioinformatics resources you need?

□ Colleagues	□ Workshops

□ World Wide Web	□ Literature
□ Conference	Others (Please specify:) □

7. Please indicate how you access the online bioinformatics resources: (multiple choice)

□ WWW-interface	□ FTP
□ Email	□ Network client via TCP/IP
□ Personal correspondence	Others (Please specify:) □

8. Please check all the types of analysis you routinely perform: (multiple choice)

□ Sequence similarity searching	□ Other DNA analysis including translation
□ Functional motif searching	□ Primer design
□ Sequence retrieval	□ ORF analysis
□ Multiple sequence alignment	□ Literature retrieval
□ Restriction mapping	□ Protein analysis
□ Secondary and tertiary structure prediction	□ Sequence assembly
□ Phylogenetic analysis	Others (Please specify:) □

9.When doing information search, such as a motif search, homology alignment, how many methods do you tend to use:

○ Only 1 ○ A few (2-4) ○ many

10. How easy do you find it to keep up to date with current bioinformatics sources?

○ Extremely difficult ○ Difficult ○ Not too difficult

○ Fairly easy ○ Extremely easy

11. In general, how do you rate the following aspects of the online biological database you currently use? Use the scale from 1 (Very poor) to 9 (Excellent).

Presentation	○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9
Authority	○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9
Result reliability	○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9

Speed	○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9

12. Which of the following types of queries do you frequently ask of bioinformatics resources? (check all that apply)

□ Web-based query	□ Unix based SQL command line query
□ Email	□ Run an automated process (e.g. Perl script)
Others (Please specify:) □	

13. In general, how do you rate the ease of formulating a query?

○ Extremely difficult ○ Difficult ○ Not too difficult

○ Fairly easy ○ Extremely easy

14. How do you rate the interface in terms of its interactivity when you make your query? For example, is it easy to alter your experiment or parameters to change query result?

(poor) ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9

(excellent)

15. How easy is it for you to express your queries over many information sources at once?

(very difficult) ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7

○ 8 ○ 9 (very easy)

Section III---Future needs

16. How important is online bioinformatics resources in the advancement of your biological knowledge ?

(not important) ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7

○ 8 ○ 9 (very important)

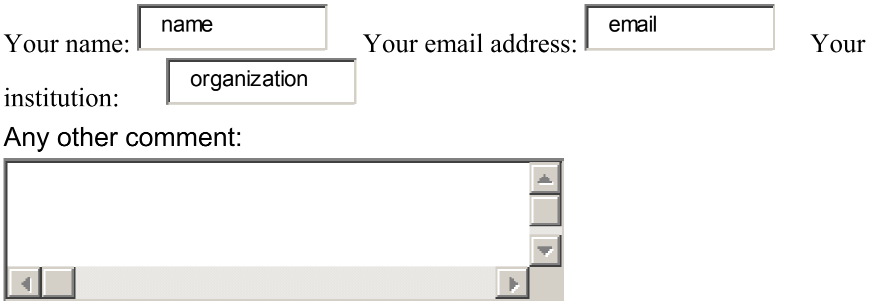
17. How do you anticipate your personal use of online bioinformatics resources?

○ increase ○ decrease ○ stay the same

18. Are you interested in attending workshops or taking class in Bioinformatics?

○ yes ○ no

Section IV---Optional personal information (you can skip this part)



Your name: name Your email address: email Your institution: organization

Any other comment:

Appendix II: Attachment for AA-IRB Proposal Form

Project description:

A web survey will be conducted on how biologists access and use online bioinformatics resources for genome related research. The survey will be distributed via email and the survey results will be saved into the database directly. Dr. Gary Marchionini is my advisor for this research project.

Participants:
We expect there will be approximately 50 participants. The survey will be sent to the biologists selected from the members of The Arabidopsi*s* Information Resources (TAIR) which is a major genome research society. The Arabidopsis is among several model organisms that is sequenced or being sequenced such as Yeast, E. Coli, Drosophila etc. The inclusion criteria is (1) that they are working biologists include faculty, graduates, either in academics or industries and (2) that they are working on genome related projects.

Are participants at risk:
The participants are not at risk.

Are illegal activities involved?
There are no illegal activities involved in this study.

Is deception involved?
No deception is involved in this study.

Prior Consent.
Implicit prior consent will be attained. It is assumed that if a person completes the survey, he or she has consented to participate.

Describe security procedures for privacy and confidentiality:
The information the participants provide will be stored in password-protected database. It will be used for only research purpose. We will make every effort we can to protect this information. No results will identify individuals in anyway.

Appendix III: Invitation letter

Dear Ms./Mr.,

My name is Dihui Lu, and I am a graduate student in School of Information and Library Science in University of North Carolina at Chapel Hill. I am writing to invite you to join our survey on how biologists access and use online bioinformatics resources for genome related research. This project is part of my work toward MS degree in Information Science and Dr. Gary Marchionini is my advisor. We got your contact information from the TAIR website (The Arabidopsis Information Resource, http://www.arabidopsis.org). The ultimate goal of this research is to better understand how biologists access and use online bioinformatics resources such as databases, software. Your willingness to share your opinion will be valuable not only for our research, but also for future development of bioinformatics applications to serve the biologists.

The survey is available at http://kiwi.ils.unc.edu/projects/bioinfo/survey.cfm. It will take approximate 10 minutes for you to complete this questionnaire. Your participation is completely voluntary. We guarantee that all information gathered from this questionnaire will be anonymous and will be kept in password-protected database. You can review the survey and decide not to respond to any reason and you may also decide not to respond to certain questions. Your submission of the questionnaire form will be taken as indication of your consent to participate in this project.

If you have any questions or comments, please feel free to contact me (lud@ils.unc.edu) or Dr. Gary Marchionini (march@ils.unc.edu). If you have any concerns about your rights in this study, please contact the Chair of the AA-IRB (Academic Affairs Institutional Review Board) of UNC-CH, Dr. Barbara Davis Goldman at 919-962-7761 or email to: aa-irb@unc.edu. Thank you in advance for your participation in our project.

Sincerely,

Dihui Lu
Lud@ils.unc.edu
Tel: 919-914-7562
School of Library and Information Science
University of North Carolina at Chapel Hill