Letting the Genes Out of the Bottle: Making the Most of Free Bioinformatics Information

Don MacMillan
University of Calgary, Canada
macmilld@ucalgary.ca

Abstract
Academic libraries can add quality to their services by going beyond standard bibliographic resources. In order to maintain relevance to the teaching and research missions of our institutions, libraries need to widen their range of information sources to include emerging data repositories. Bioinformatics provides a clear example of a discipline rich in freely available information that students and faculty can use to create new knowledge. Adding instruction in using key genetics databases to an existing suite of information literacy sessions allowed this librarian to expand his own subject knowledge, develop a productive collaboration with a faculty member and provide a high-quality, innovative learning experience for students.

This session will place bioinformatics databases such as OMIM (Online Mendelian Inheritance in Man), BLAST (Basic Local Alignment Search Tool), and ExPASy’s ProtParam (Protein Parameter) within the context of library reference and instruction. Critical aspects of each resource and the joys and pitfalls of learning and using them will be outlined as examples for effectively incorporating non-bibliographic resources into research assignments. Participants will be invited to contribute their own experiences with using similar tools.

Keywords: Bioinformatics, genetics, information literacy

Introduction
Academic libraries can add quality to their services by going beyond standard bibliographic resources. In order to maintain relevance to the teaching and research missions of our institutions, libraries need to widen their range of information sources to include emerging data repositories. Bioinformatics provides a clear example of a discipline rich in freely available information that students and faculty can use to create new knowledge. Adding instruction in using key genetics databases to an existing suite of information literacy sessions allowed this librarian to expand his own subject knowledge, develop a productive collaboration with a faculty member and provide a high-quality, innovative learning experience for students.

Bioinformatics databases such as OMIM (Online Mendelian Inheritance in Man), BLAST (Basic Local Alignment Search Tool), and ExPASy’s ProtParam (Protein Parameter) provide critical access to genetic data used by researchers across the life sciences. Bioinformatics data pose both a challenge and an opportunity to librarians. The challenge arises because while some of our knowledge of how information works transfers well to these new sources, the nature of the information they contain requires significant learning and practice on our part in order to help clients and students use them effectively. However, the opportunities these resources present to provide resources for research, to retain professional credibility and relevance, and to expand our own skill sets are well worth the time and effort required to master them.

Before describing the case study of an instruction session that introduced students to using non-bibliographic data, a little background is in order. The information environment is changing and the end-user now has far greater access to both bibliographic and non-bibliographic material than ever before. The rate at which discoveries are made often outstrips the formal publication process, which has led to databases themselves becoming prime sources of information as scientists load raw data directly into them. Becoming familiar with tools that make information public ahead of the usual channels of articles and conference papers is imperative if librarians are to retain their place in information management and dissemination. In the social sciences, large data sets are increasingly available for users to manipulate; in technology patents often serve as the only formal publication of innovation, and in genetics, the bioinformatics data sets are at the cutting edge of research and discovery.

While some might argue that the responsibility for teaching users to make the most of these
information resources rests with discipline faculty, many authors have argued that librarians will have to develop some degree of fluency with non-traditional information sources) and increase their instruction activities in these areas in order to remain relevant [Bowden & DiBenedetto, 2001; Brown, 2005; Chiang, 2004]. In an academic library setting, teaching the use of these advanced resources can be an excellent marketing tool, encouraging faculty to bring students back to the library in senior courses. This provides an opportunity to review prior bibliographic skills as well as introduce new resources. As librarians we have the training and experience not only to select the tools that can best help our users but also to make sense of complex interfaces, and to teach information skills effectively. Bioinformatics data provide a new platform for our expertise, one where we are uniquely able to contribute to research and education.

Bioinformatics

The discipline of bioinformatics combines the fields of computer sciences, mathematics, and the biological sciences to manage and make accessible the vast amounts of data generated by gene sequencing projects. In addition to changing how biological research is conducted, Miskowski, Howard, Abler & Grunwald described bioinformatics as changing the types of questions that can be asked and increasing the rate at which knowledge is obtained [2007]. The American Society for Biochemistry and Molecular Biology has recommended that biochemistry and molecular biology undergraduate curriculums include the use of computer databases and bioinformatics as a core competency before graduation [Bednarski, Elgin & Pakrasi, 2005; Boyle, 2004; Voet, 2003]. Andrew Feig and Evelyn Jabri go further to state "supplementing the undergraduate biochemistry curriculum with data-mining exercises is an ideal way to expose the students to the common databases and tools that take advantage of this vast repository of biochemical information" [2002]. Furthermore, as Feig and Jabri also point out: Bioinformatics tools are freely available through the Internet, they are relatively easy to incorporate into the curriculum and there is no need to purchase expensive hardware or software, thereby eliminating common barriers to integrating a new information source into library services.

There is extensive literature on the benefits of incorporating bioinformatics within library work. The July 2006 issue of the Journal of the Medical Library Association was dedicated to bioinformatics and the role of medical libraries. This issue includes a number of examples of the integration of bioinformatics into library information literacy sessions for graduate or medical students. There is widespread recognition that bioinformatics databases should be included in the repertoire of life and heath science librarians in order to retain our role in supporting the work of molecular biologists [Brown, 2005; Geer, 2006; Tennant, 2002].

There may be some social barriers to integrating bioinformatics databases into library services. Scientists may think that librarians have insufficient knowledge of genetic information to provide assistance, or that the databases are too advanced or complex for their students to understand. As with many areas of librarianship, the key to overcoming these challenges is collaboration. The librarian can contribute expertise in how databases function, and effective pedagogies for guiding students in using information resources. The discipline expert brings the knowledge of how the databases are used by scientists, and which kinds of data are extracted to provide what kinds of information. Working together, they can develop an authentic learning experience for students. As is evident in the case study, the more closely the assignments model the attainment of specific 'real world' information goals, the more readily the students see the application of assignments to their current and future work.

The National Center for Biotechnology Information (NCBI) is the source for most, though not all bioinformatics databases, providing access to more than 30 databases. GenBank serves as the core resource, a large annotated collection of nucleotide and amino acid sequence data that includes over 83 million sequence records or 86 billion base pairs of nucleotides. Using this data, scientists can, for example, compare one or more gene sequences to see if they share a common ancestor. GenBank includes human sequence data produced by the human genome project and that of over 133,000 other species [Rapp and Wheeler, 2005]. The Entrez interface familiar to PubMed users, provides the search functions for a resource that incorporates data from all the major NCBI databases on DNA and protein sequences, gene expression, genetically inherited diseases, bibliographic databases such as PubMed, and genome maps. Databases such as OMIM (Online Mendelian Inheritance in Man) and BLAST (Basic Local Alignment Search Tool), used in the case study, link into GenBank, providing various entry points to the sequence data that serve specific research needs. These tools enable scientists and researchers to locate, extract, assemble, and analyze data from which they may develop critical insights in the health and life sciences. These databases, OMIM and BLAST, and one other, ExPASy’s ProtParam (Protein Parameter) formed the basis of the information literacy session.
and assignment in the case study. As background for the Case Study, they are introduced briefly below.

**OMIM - Online Mendelian Inheritance in Man**

OMIM ([Online Mendelian Inheritance in Man](http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim)) is a catalogue of human genes and genetics disorders created at Johns Hopkins University. At present there are over 19,000 entries in OMIM, used by researchers and medical practitioners to advance the knowledge and treatment of genetic disorders. Each record contains textual information and references about the gene(s) believed to be responsible for a specific disorder including its location, history, and inheritance pattern as well as clinical features of the disease and links to references available through PubMed. Users can search with keywords or by using the Limits or Preview/Index options to narrow searches by Chromosome Number or a known gene sequence. Readers may wish to refer to the site at: [http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim](http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim).

**BLAST**

BLAST ([Basic Local Alignment Search Tool](http://www.ncbi.nlm.nih.gov/blast/Blast.cgi)) is another NCBI computational tool that enables users to quickly compare gene sequence data. A researcher or student can use the BLAST program to compare one gene sequence against all others in a database to identify similar ones. BLAST provides a score based on the degree of similarity between the sequences, allowing you to measure the degree of similarity. For example, someone who has cloned and sequenced a human gene may want to see if this gene is similar to any other genes in the genome or to genes in other species. Readers may wish to refer to the site at: [http://www.ncbi.nlm.nih.gov/blast/Blast.cgi](http://www.ncbi.nlm.nih.gov/blast/Blast.cgi).

**ExPASy-ProtParam**

ExPASy-ProtParam is a tool that enables users to calculate various physical and chemical parameters for a particular protein stored at the Swiss Protein Knowledgebase. These parameters include a protein's molecular weight and amino acid composition. ProtParam is maintained by the Swiss Institute for Bioinformatics, and is not directly linked to the NCBI system. However, the data it provides on proteins complements information available in OMIM and BLAST and is often more current and complete. Readers may wish to refer to the site at: [http://ca.expasy.org/tools/protparam.html](http://ca.expasy.org/tools/protparam.html).

**Case Study - Biology 311-Introduction to Genetics**

In 2007 this librarian worked with a faculty member to introduce students in Biology 311, Principles of Genetics, to key tools used by molecular biologists; PubMed, OMIM, ExPASy’s ProtParam and BLAST. Patent information was also incorporated in the assignment, but will not be described in this paper. The session was offered in November and has been revised and repeated in November 2008. Students were required to locate data using these tools to complete an in-class exercise, and were also expected to use the resources in their major poster presentation worth a percentage of the overall laboratory mark. The ExPASy-ProtParam database was added to the curriculum in 2008 so that students could more easily answer specific questions about the physical and chemical parameters (e.g. calculated molecular weight and number of amino acids) for the protein related to their chosen gene. The instructor and I developed the workshop to lead students through complex databases in a series of steps that reflected how molecular biologists would use the tools to research a particular disorder. By actually using the tools to answer questions, students developed an understanding of the vast amount of data available and the types of questions that could be answered using bioinformatics databases. While most similar classes described in the literature are designed for graduate students, this session was developed for 2nd or 3rd year students. However, with a well-designed assignment where learning was supported by clear directions and hands-on use of the tools, these students were able to find their way through very advanced information resources.

Prior to the library class, students selected a genetic disorder from the "Genes and Diseases" web site on NCBI at [http://www.ncbi.nlm.nih.gov/books/bookres.fcgi/qnd/tocstatic.html](http://www.ncbi.nlm.nih.gov/books/bookres.fcgi/qnd/tocstatic.html). The session took place in the library’s computer lab, allowing demonstration and hands-on work with the resources. I began with an introduction to PubMed with which students were required to locate an article about ‘their’ disorder. This reviewed prior knowledge of other bibliographic databases such as BIOSIS and Web of Science and served as a bridge using the ENTREZ common interface to OMIM and BLAST. For the next stage of the class, an OMIM search, students had to look for their topic and select the record for the gene thought to be responsible for the disorder. Using the information from OMIM they had to answer a series of questions, which required them to use different parts of the record. Students were given time to do this in class so that the instructor, lab assistants and I could answer questions and ensure they understood the structure of the information. One particular item, identifying a protein
sequence, provided the information for the next stage of the exercise, using ExPASy’s ProtParam.

Students then copied the protein sequence information found in OMIM and subsequently pasted this information into ExPASy’s ProtParam search box. This search produced the calculated molecular weight and number of amino acids for their protein. Students then used the same protein sequence they located through OMIM to conduct a BLAST Search. Again, they had to answer specific questions that required them to use the database as a researcher would, comparing data across various species.

Here is a condensed summary of the example used for Biology 311:

- Search “Cystic Fibrosis”.
- Select record “602421”
- In the upper right corner, select -> Links
- Select -> Gene
- Answer questions on your chosen gene
- Under Links on the right side, select -> Protein
- From the resulting list, select the first record noting [Homo sapiens] - NP_000483
- Copy the Protein sequence from the bottom of the NP_000483 record
- Go to ExPASy-ProtParam database - http://ca.expasy.org/tools/protparam.html
- Paste the protein sequence found earlier in the NP_000483 record
- Select -> Compute Parameters
- Note the number of amino acids and Calculated Molecular Weight for your protein
- Using the NP_000483 record found earlier, copy the GI number (Sequence Number), gi|90421313|
- Under Basic BLAST, select -> Protein Blast
- Enter the GI number copied from the OMIM record or use protein sequence copied earlier
- BLAST Search takes a few minutes to run-wait for results page
- For the full lab exercise, please see: http://library1.ucalgary.ca/uploaded_files/subject_documents/Library/57829/BIOL311labmanaulabmanual08.pdf

Throughout the sessions the students were engaged by the exercises and intrigued by the information resources. After they had completed the structured assignment, several explored other aspects of the data, search options and help screens. Like a map to an unknown city, the structured exercise provided students with clear paths to specific information; once they were familiar with the main thoroughfares, they felt much more comfortable exploring side roads and back alleys. Their ability to use the information was evident in the poster presentations that formed the major assignment for the course. Here students clearly demonstrated they had learned more about genetics than they could have gained just from the lectures and textbook. More than that, they gained an understanding of how molecular biologists use information to develop knowledge.

The instructor and I met several times after the session to review whether it had accomplished the instructor’s goals and how it might be improved. Noting that students had encountered difficulties deriving basic data about proteins from OMIM where the required data was not always available or precise, we explored other resources and the instructor decided to add ExPASy’s ProtParam to the exercise. Having seen how well the students were able to work through two bioinformatics resources, I was not concerned about their ability to handle a third. In the second iteration of the library exercise in November 2008, students indeed had no difficulty switching out of the NCBI databases to complete that portion of the assignment and were able to spend more time using the data and less time trying to find or verify incomplete information. Again, in the poster presentations, which dealt with the same
topics as their library exercise, it was clear that students had adopted these new resources into their workflow and were comfortable using the data.

Assessment
Student satisfaction with the library portion of the library session was measured in part using “getfast”, a free web-based assessment tool that allows students to provide anonymous feedback on a class or instructor. Feedback was also collected at the conclusion of the student poster presentations. Most students felt the library sessions and presentations were a meaningful learning experiences that introduced them to the tools used by researchers in the “real world”. Students also felt that by organizing and presenting their work from the various databases they gained a deeper understanding of these resources than would be the case with an exam or laboratory report.

Sample student comments from the GetFast survey included:

- “I learned how to use PubMed How to find a specific protein How to find the coding sequence for the specific protein”.
- “OMIM is quite complicated but the instruction provided really helped to make this website useful”.
- “I have used Patents before (at Mount Royal College) and it was great to have an in-depth review of the hits that one generates when using Google Patents”.
- “If I use OMIM again in the future it will be for any extra genomic information that I need about a disease”.

Readers may wish to refer to the site at: http://www.getfast.ca/

Based on my experience in integrating bioinformatics databases into information literacy instruction, I offer the following suggestions which may be useful for those wishing to incorporate advanced, non-bibliographic materials into their library work.

Tips for integrating bioinformatics databases in library instruction

1) Bioinformatics and other non-bibliographic resources are complex and generally contain more information than students need in an introduction. By initially focusing on the most useful aspects of resources, you can reduce frustration and encourage later exploration of other aspects of the resources.

2) Break the exercise down into sequential clear steps that take students where they need to go to complete the focused assignment. Success breeds confidence, and students who can successfully complete one task in a database are likely to use it again.

3) As far as possible, develop an exercise that that replicates the way experts use the resources. Using a resource to ask ‘real’ questions lends credibility to the exercise. Also – make sure that students who effectively incorporate material from the new resource in their later work are rewarded for it, through rubrics, assignment specifications, etc.

4) Ask targeted questions for specific data to encourage students to read the screens carefully, extract what they need, and use built in capabilities of the resource to manipulate the data – e.g. calculate the molecular weight of a protein, map trends, graph data, etc.

5) Check the terms that you and the students will be using to search the resources in advance – in our case, the data for some diseases are easier to extract than others, and the amount of information available for each disorder varies considerably.

6) Provide time for hands-on practice with readily available assistance from the librarian and the instructor. Questions will arise with searching and interpreting the data that both experts may be required to help with.

Conclusion
The sessions on using bioinformatics information provided considerable benefits for the librarian, the faculty, the students, and ultimately the institution. In preparation for the class, I broadened and deepened my skill set, and it was enlightening to stand in the shoes of the student once more, learning something new with all the excitement, stumbling, frustration, and accomplishment that entails. Developing closer partnerships with faculty has provided insights into how researchers in biology approach information, improving my understanding of their collection needs, and the information
literacy requirements of their students. These collaborations are rewarding partnerships; I bring an expertise in information and an understanding of the novice researcher's perspective, making sure that explanations are clear, that assignments are actually do-able and that the work builds on and reaffirms existing literacy skills. Classes in advanced information resources benefit other library personnel as well as I often draw on other staff and librarians to assist with large classes; these librarians and staff also learn about the resources during the class and are then better able to assist students at the reference desk.

For the professor, working with a librarian may provide new ways of using known resources or exposure to entirely new data sources. By situating instruction in bioinformatics in the context of other information resources, the course may strengthen links to prior learning. Finally, through teaching students not just how to 'do' molecular biology, but rather how to be molecular biologists using the same tools to answer the same questions, we can invite them into a community of practice and better prepare them for work in the field.

The students welcome the opportunity to work at the cutting edge of science - seeing the raw material of new discoveries in bioinformatics data. Indeed students may be more apt to see the relevance of these resources than of texts and journal articles, which can often distance the students from the possibility of discovering something new that likely drew them to the sciences in the first place. Having seen this data they may develop a greater appreciation for the articles which make sense of it, and the patient work of contextualizing and communicating innovation.

The resources are there for us to use; the benefits are clear - Dive in!

Acknowledgements

I would like to thank Dr. Isabelle Barrette-Ng and the students in Biology 311 for letting me participate in this initiative.

References


