Towards A Deeper Understanding: Linking Mendelian Genetics With Molecular Genetics Using Web-Based Bioinformatics Tools

Isabelle H. Barrette-Ng[1], Don MacMillan[2], and Dave Hansen[1]

[1] Department of Biological Sciences, University of Calgary, 2500 University Drive N.W., Calgary, Alberta, CANADA, T2N 1N4
mibarret@ucalgary.ca
dhansen@ucalgary.ca

[2] MacKimmie Library, University of Calgary, 2500 University Drive N.W., Calgary, Alberta, CANADA, T2N 1N4
macmilld@ucalgary.ca

Abstract

The traditional model of teaching introductory undergraduate genetics begins with a description of classical Mendelian genetics followed by an abrupt transition to topics in molecular genetics. We re-evaluated this model and asked whether the teaching of Mendelian and molecular genetics could be improved by introducing an inquiry-based laboratory exercise in which pairs of students investigated a genetically-inheritable disease using basic bioinformatics tools. The implementation of this exercise in a large (>500 students), second-year undergraduate, introductory genetics course for biology majors led to a high level of student satisfaction and a more integrated introduction to classical and molecular genetics.

Introduction

Both historically and pedagogically, the basic principles of Mendelian genetics have provided a sound and extremely powerful framework for understanding patterns of inheritance at a phenomenological, organismal level. As a result, most introductory genetics courses begin with a thorough explanation of the principles of Mendelian genetics (Haffie *et al.*, 2000; Sheley and Mertens, 1990). Typically, students are only introduced to topics on biochemical and molecular genetics after a thorough examination of Mendelian principles. The later introduction of these topics is usually meant to help provide students with a molecular explanation for the basic mechanisms underlying the core concepts of Mendelian genetics. Although this framework has generally worked well and is commonly used by many introductory genetics courses and textbooks, this approach does have important pedagogical shortcomings, some of which have previously been investigated and discussed (Bednarski *et al.*, 2005; Bowling *et al.*, 2007; Dymond *et al.*, 2009; Holtzclaw *et al.*, 2006).

The most important problem that we have encountered in teaching introductory genetics at the undergraduate level has been the tendency for many students to compartmentalize the sections of the course dealing with Mendelian genetics and the sections dealing with molecular genetics. Although experts can clearly see the relationships between these two areas, many students have more difficulty recognizing these relationships. We found that many students overemphasize the distinctions between the concepts and approaches taken in Mendelian genetics versus those taken in molecular genetics. For example, students usually grasp the concepts of dominance and recessiveness within the classical Mendelian framework, but often fail to see how variations in DNA sequence underlie the nature of dominant and recessive alleles at the molecular level. Linking such concepts between Mendelian and molecular genetics is essential to encourage a broader understanding of genetics; linking these concepts is a very important pedagogical challenge facing all introductory genetics instructors.

To help students better understand the connection between classical and molecular genetics in our introductory genetics course, we designed and implemented a novel, inquiry-based, experiential learning laboratory exercise. In this exercise, students use modern, web-based bioinformatics tools and databases to link concepts in Mendelian genetics with those in molecular genetics. We specifically designed this laboratory exercise using these tools and databases to provide students with an inquiry-based learning experience. The importance of integrating bioinformatics tools and databases into the undergraduate science curriculum has already been advocated by many, including the National Science Foundation and the American Society for Biochemistry and Molecular Biology (Boyle, 2004; Feig and Jabri, 2002; Miskowski *et al.*, 2007; Voet *et al.*, 2003).

Bioinformatics tools and databases were particularly attractive to us in the development of the inquiry-based exercise for two main reasons. First, the databases provide vast amounts of molecular and Mendelian genetic data obtained through cutting-edge technologies, in response to current research questions. Second, the wide array of freely available bioinformatics tools engages students to become active participants in the study of current research questions. These tools and databases allow students to select a topic of interest, and step into the shoes of a geneticist, where they are asked to apply their theoretical knowledge in solving a real-life question. Students select a human genetically-inheritable disease and research both Mendelian and molecular aspects of the

disease. It has been our experience, as well as that of others (DeHaan, 2005), that students develop deeper and more substantive links between different subject areas when they are encouraged to develop these links in a way that is personally meaningful. Because many students have personal experience with genetically-inheritable diseases, presenting students with the option of choosing a specific disease as the topic of their research project promotes student engagement and stimulates active learning.

Several important practical considerations needed to be overcome for us to implement this inquiry-based exercise. Most importantly, we needed to implement this exercise in a way that could accommodate the large numbers of students (>500) enrolled in our introductory genetics course for both majors and non-majors in biology, at a publicly-funded university with limited resources. To achieve this, it was critical for us to access the expertise in information literacy, as well as the extensive computing infrastructure provided by our university's academic librarians. Throughout the exercise, students are guided through a self-motivated inquiry process by academic librarians, instructors, and graduate teaching assistants. Students appear to welcome the opportunity to work at the cutting edge of science – seeing the raw material of new discoveries in bioinformatics data. By analyzing these data, students can better appreciate the relevance of Mendelian genetics and molecular genetics to understanding common human diseases. From Mendel to molecules to disease, our inquiry-based exercise challenges students to apply their understanding of genetics in combination with cutting-edge bioinformatics tools to explain how many diseases are inherited.

Student Outline

The text below represents the description of the month-long exercise that is provided to students at the beginning of the semester. In the first week of the exercise, students are presented with a list of 80 genetically heritable diseases and are asked to select, in groups of two, a disease of interest. The actual list of diseases used for our class is presented in APPENDIX A.

In the second week, students perform searches of peer-reviewed scientific publications using the PubMed database and gene-based therapies using the Google Patents database to uncover the breadth of basic information and gene-based applications related to their chosen disease. Specifically, students select two review articles and a patent. The exercise was structured to encourage students to use the basic genetics concepts they had recently learned through the lecture and laboratory components of the course to allow them to develop a basic understanding of their chosen disease.

In the third week, students employ an array of bioinformatics tools to discover the molecular basis and inheritance pattern of the heritable disease chosen for study. Interactive training sessions are conducted by a combination of librarians and teaching assistants to show the students how to effectively use these tools to investigate their specific diseases.

In the final week, each pair of students organizes and presents the results of their research to the other students in their laboratory section, as well as to a graduate teaching assistant. Each pair of students prepares a poster in which they present (1) basic background information on their chosen disease, (2) the classical Mendelian pattern of inheritance (e.g., autosomal or sex-linked, dominant or recessive), (3) molecular information on the genetic basis of the disease (e.g., how does the molecular information explain the Mendelian pattern of inheritance?), and (4) practical, gene-based therapies described in the patent literature.

What causes human genetically-inheritable diseases?

Over the next four weeks, you will have the opportunity to study, using various bioinformatics search tools and databases, a human genetically-inheritable disease that you will select. Using your chosen disease, you will be introduced to various databases and search tools that are routinely used by most researchers. Working with your laboratory partner, you will collect various data relating to your chosen disease and then summarize your findings in a research poster, which will be presented to your laboratory section. The schedule for the four weeks is shown in **Table 1**.

Table 1. Laboratory exercise schedule

Week	Activity
1	Selection of human genetically-inheritable disease
2	Selection of peer-reviewed journal articles and patent documents relating to your chosen topic
3	Use of bioinformatics search tools and databases to study the Mendelian and molecular genetics aspects of your chosen topic
4	Preparation and presentation of research poster

Week 1: Choosing a disease for study

The very first step in this project is choosing a disease for which all the necessary information for this laboratory is available. You will need to choose a disease you are interested in, capable of making a presentation on, and for which a causative gene has been identified and isolated in humans.

An ideal starting point is "Genes and Diseases" list on the NCBI website. The link is as follows:

http://www.ncbi.nlm.nih.gov/-> [Left Frame] "Literature Databases"-> [Left Frame] "Genes and Diseases"

This compilation is a new addition to the NCBI database, and the number of diseases listed in rapidly growing. Be careful – many of these diseases can be caused by a variety of mutations in several genes, combinations of mutations, or wide-spread deletions. It is essential you pick a single gene involved in causing a particular disease.

In researching your chosen disease, ensure that the gene(s) responsible for the disease have been identified (unlike "Asthma", for example), and have been cloned in humans. Exceptions to either will almost always be noted in the brief summary provided on the Genes and Diseases website. However, to be safe, it's a good idea to search for your selected gene in a database such as OMIM. The link is as follows:
http://www.ncbi.nlm.nih.gov/ ->[Top Frame] "OMIM" ->Search "yourgene" ->[Right] "Links->Gene" AND [Right] "Links->Protein"

Once you have found the gene and protein in *"Homo sapiens"*, you can be sure that all of the information you need to complete the assignment is available, and breathe a sigh of relief.

In order to help you select your disease, we have compiled a list of diseases that meet these criteria. The list is attached to this document as APPENDIX A.

Please consult with your laboratory partner and select a few diseases of interest (including selected gene). **Topic selection will be on a first come, first served basis. Each disease may only be selected once per laboratory section.** Please let your G.T.A. know of your selection during your laboratory session.

Week 2: How to find peer-reviewed journal articles and patents relating to your chosen genetically-inheritable disease

In this portion of the exercise, you will learn how to locate peer-reviewed journal articles and patents that relate to your chosen genetically-inheritable disease.

(A) Peer-reviewed journal articles

As you probably already know, peer-reviewed journal articles are only published after the article has been reviewed by a number of research scientists who are currently involved in the same or similar research area as that described in the article. To be accepted for publication, the article must convince the reviewers that the results described are scientifically sound, convincing, new and, for many journals, of interest to the broad scientific community. Consequently, the process of peer review helps ensure that only scientifically sound and reproducible results are disseminated to the scientific community.

Peer-reviewed journal articles usually take the form of either a research article or a review article. In a research article, new results are reported to the scientific community. In a review article, results that have been published in a specific research area over a period of time are reviewed and summarized. This type of article also attempts to identify any questions that remain to be answered within that field, as well as introduce models and/or mechanisms based on available results.

Each year, many thousands of peer-reviewed journal articles are published in a wide variety of journals. Today, the task of finding a peer-reviewed journal article on a specific topic is much easier due to the advent of publicly available databases. PubMed is currently one of the most frequently used databases. In fact, in May of 2007, the National Centre for Biotechnology Information (NCBI) reported that there were 78 835 000 searches performed in that month alone! PubMed is freely available through the NCBI website, which can be found at the following address: http://www.ncbi.nlm.nih.gov.

The NCBI was first established on November 4, 1988, as a division of the National Library of Medicine (NLM) at the National Institutes of Health (NIH). It maintains an extremely useful collection of databases. As a publicly-funded U.S. governmental unit, it is freely accessible to the public. You should notice that across the top of the NCBI website, there are buttons labeled “PubMed”, “All Databases”, “BLAST”, “OMIM”, “Books”, “TaxBrowser” and “Structure”. These are some of the most common search tools or databases used by researchers today. Over the next two weeks, you will be introduced to a few of these.

As of November 13, 2008, PubMed includes 18,463,308 citations from MEDLINE and other life science indexes for biomedical journals that date back to the 1950s. This database is extremely useful as it easily and quickly allows you to locate published information on your selected topic.

Additionally, PubMed includes links to full-text articles and other related resources. The University of Calgary has electronic subscriptions to many journals, which allows you to quickly download a copy of an article.

Through this exercise, you will gain experience in conducting searches in PubMed. As you will see, searches can be conducted using many different parameters such as keywords, names of authors, names of journals, and names of articles. During the session at the library, you will be shown how to perform these searches easily and effectively using specific parameters.

At the end of your session at the library, you and your laboratory partner must each have chosen one peer-reviewed review article relating to your chosen disease. You will be using both of these review articles as background information when making your poster. **Before leaving the instructional session at the library, you must give your graduate teaching assistant the reference of your review articles using the Harvard style of referencing.**

(B) Patents

(i) What is a patent?

A patent is protection that is granted for an invention. This protection allows one to exclude others from making, using or selling an invention from the day the patent is granted to a maximum of twenty years after the day on which the patent application was submitted to a governmental patent granting agency. In exchange for the grant of a patent, the inventor (or patentee, who is the owner of the patent) must provide a detailed description of the invention so that all can benefit from this advance in technology and knowledge.

(ii) What is an invention?

An invention is considered to be any new or useful object, process, machine, composition of matter or manufacture. Any improvements to existing objects, processes, machines, compositions of matter or manufacture are also considered to be inventions.

(iii) How are patents granted and where are they valid?

In Canada, patent protection is granted by the Canadian Intellectual Property Office under the guise of the Canadian *Patent Act*. You can find more information on patents in Canada on the following website: http://strategis.ic.gc.ca/sc_mrksv/cipo/welcome/welcom-e.html. This website also contains links to a few tutorials that describe how patents may be obtained (http://strategis.ic.gc.ca/sc_mrksv/cipo/patents/pat_gd_protect-e.html#sec1) as well as how patents should be written (http://strategis.ic.gc.ca/sc_mrksv/cipo/patents/e-filing/menu.htm).

To be granted a patent, an inventor must demonstrate to the granting agency that the invention is new (first in the world), useful (functional and operative), and not obvious to others working in the same area (*i.e.*, people should have the reaction of "I wish I had thought of that!"). Only once all three requirements have been demonstrated will the patent be granted.

A patent granted in one country is only valid in the country it was issued. For example, a patent granted in Canada would not prevent others in another country from using the invention. Consequently, most inventors will seek patent protection concurrently in many different countries.

(iv) What does a patent look like?

Patents are usually composed of four main parts:

1. abstract;
2. description;
3. claims; and
4. drawings

The **abstract** provides a brief summary of the invention, much like the abstract of a journal article.

The **description** provides a detailed description of the invention, so that after the expiration of the patent, anyone interested can reproduce the invention.

You can think of **claims** as fences. The claims establish the area that has become the monopoly of the inventor for a certain period of time. They tell any competitors that the subject-matter described in the claims or contained within those "fences" belongs solely to the inventor for the set period of time.

Drawings are not always present in every patent. Sometimes, they are used to provide a better description of the claimed invention.

(v) Patents and the biotech industry

Due to increasing research and development costs, many pharmaceutical companies are now seeking patent protection for various research projects. More and more, pharmaceutical companies, as well as university researchers, are increasingly seeking patent protection to recoup research costs incurred over the many years it took to identify the gene(s) responsible for specific diseases. A patent granted for a specific gene will allow the patent holder to exclude others from working on this gene for a period of time. In this way, the holder can continue researching the gene with the goal of finding a possible treatment to the disease without fear that competitors may reach this goal in a shorter amount of time.

Over the last few decades, the number of patents describing biotechnological inventions has risen dramatically. Patents can now be granted for specific genes, proteins, protein structures, laboratory techniques and compounds useful in the treatment of various diseases. Due to the rapid increase in biotech patents, patent databases have now become an increasingly richer source of biotechnological knowledge. Many researchers now look to patent databases, as well as to databases like PubMed, in order to locate information on specific topics.

(vi) How can you find a patent?

There are now many publicly available patent databases. Because a patent granted in one country is only valid in that country, most countries have their own patent database. In Canada, the patent database can be found by using the following link: http://patents1.ic.gc.ca/intro-e.html. In the U.S., the patent database can be found by using the following link: http://www.uspto.gov/patft/index.html. Databases like esp@net provide links to patents from

multiple countries. However, some of these databases are sometimes difficult to search or require specialized software to view the patent documents.

Recently, Google launched a search portal that simplifies searching for patents issued by the U.S. Patent and Trademark Office (USPTO), which contains over 7 million patents. Google patents may be accessed using the following link: http://www.google.com/patents. This search tool allows you to search patents using many different parameters such as inventor name, patent number, and topic.

(vii) Your exercise

In this exercise, you will use Google Patents to find a patent that relates to your chosen genetically-inheritable disease. During the session at the library, you will be shown how to perform searches in Google Patents. Each genetically-inheritable disease that was listed in APPENDIX A should have a patent that discusses it, although sometimes in a remote fashion. You and your laboratory partner will together be asked to find only one patent.

After you and your laboratory partner have located one patent, you will find the following information:

- What is the title of the patent?

Your search should produce a list of records. The hyperlinked portion of each record corresponds to the title of the patent. Also, if you decide to download a PDF version of the patent, the title is also shown as entry [54] at the top of the left column on the first page of the patent.

- What is the patent number of the patent?

Once a patent is granted, a patent number is assigned. In the U.S., the patent number usually contains six to seven numbers, with all patents recently issued having patent numbers that contain seven numbers.

To find the patent number, click on the hyperlinked portion of the record of interest from the list of records you obtained. The patent number is listed in the left column of the record. If you decide to download a PDF version of the patent, the patent number is also shown as entry [11] at the top of the right column on the first page of the patent.

- What is the filing date of the patent?

The filing date represents the date on which the USPTO received the application.

To find the filing date, click on the hyperlinked portion of the record of interest from the list of records you obtained. The filing date is listed in the left column of the record. This is the simplest way to locate the filing date. Other filing dates may be listed in the PDF version of the patent, but these sometimes represent filing dates of other applications.

- What is the issue date of the patent?

The issue date represents the date on which the USPTO determined that the patent could be granted.

To find the issue date, click on the hyperlinked portion of the record of interest from the list of records you obtained. The issue date is listed in the left column of the record. If you decide to download a PDF version of the patent, the patent number is also shown as entry [45] at the top of the right column on the first page of the patent (below the patent number).

- How many claims does the patent contain?

Remember that the claims are the "fences". To find the number of claims, click on the hyperlinked portion of the record of interest from the list of records you obtained. The claims are listed in the right column of the record. The left column of the record also includes a hyperlinked "Claims" that will automatically download a PDF copy of the claims of the chosen record.

- What does the first claim of the patent say?

Here, please provide, in your own words, a summary of what the first claim states.

- How many patents cite your chosen patent?

To find how many patents cite your chosen patent, click on the hyperlinked portion of the record of interest from the list of records you obtained. The patents that cite your chosen patent are listed in the left column of the record under the heading "Referenced by".

- Do you agree that your chosen patent should have been granted by the USPTO? Why or why not?

Here, we want to know your thoughts about whether it is ethical to grant patents on specific genes. Do you think that granting patents advances scientific research and knowledge? Or, do you think that patents stifle the advancement of scientific research and knowledge?

Week 3: Bioinformatics analyses of your chosen gene

(A) What is bioinformatics?

Bioinformatics is a discipline that combines the fields of mathematics, computer sciences and biological sciences. This new discipline was born approximately a decade ago. Bioinformatics is allowing researchers to start processing the deluge of data generated through large scale projects, including the various sequencing projects. For example, *GenBank*, which is a large public annotated collection of nucleotide and amino acid sequence data (access using http://www.ncbi.nlm.nih.gov/Genbank/index.html), now includes over 61.1 million individual sequence records, which represents approximately 65.4 billion base pairs! Moreover, as of November 13, 2008, there are currently 9026 genome sequences deposited with the NCBI. Without bioinformatics, analysis of all of these data would be virtually impossible!

A wide array of computational tools have been developed that can be used to study DNA and RNA structure and function, gene expression, protein production, protein structure and evolution. These tools have had very wide applications, which include drug discovery, clinical diagnostics and agricultural biotechnology.

(B) Bioinformatics analyses on your chosen genetically-inheritable disease

(i) Locating information on your chosen disease

Earlier this semester, you were asked to pick a genetically-inheritable disease from a list (Appendix A), or from the "Genes and Diseases" link on the NCBI website (http://www.ncbi.nlm.nih.gov/books/bv.fcgi?call=bv.View..ShowTOC&rid=gnd.TOC&depth=2). You were also asked to pick a specific gene that is associated with the disease using the "Genes and Diseases" collection.

You will now locate specific information on your chosen genetically-inheritable disease and gene using the "Genes and Diseases" collection, as well as the OMIM database. OMIM stands for Online Mendelian Inheritance in Man and is available from the NCBI website using the following link: http://www.ncbi.nlm.nih.gov/sites/entrez?db=OMIM.

OMIM is a database containing annotated records of human genes and genetic diseases that is authored and edited by Dr. Victor A. McKusick and his colleagues at John Hopkins University. Each record provides textual information and references about the gene (*i.e.*, location, mapping, history, etc.) or disease (*i.e.*, clinical features of the disease, inheritance pattern, etc.), as well as links to many references available through PubMed. Currently, there are 19 947 entries in the OMIM database.

To search for information about your chosen genetically-inheritable disease, go to the OMIM website using the link given above. In the box at the very top of the page, where it states "Search OMIM for", type in the name of your chosen genetically-inheritable disease. A list of records should appear. **You must now select the correct record that pertains to your CHOSEN GENE. Many other records will likely be returned, some that may not directly pertain to your chosen topic. Only select the record that pertains to your chosen gene.**

Once you have selected the correct record, click on the hyperlinked region. This should bring you to the complete record. Note that this record contains a LOT of information. You do not have to read through all of it. However, from your quick read, you should be able to find the answers to the following set of questions:

1. What is the inheritance pattern of the disease? (i.e., dominant, recessive, autosomal, sex-linked)

2. What is the name of the gene(s) thought to be involved in causing the disease?

3. Which of these genes, if there are several, have you chosen to examine? Why did you make this selection?

4. What types of mutations have been discovered in this gene? Select one mutation to describe in your poster.

5. What is the normal function of the protein encoded by this gene? How does the mutation you selected above (in your answer to question 4) alter the function of the protein, and lead to the disease? Can you now explain the Mendelian inheritance pattern of the disease on the basis of the available molecular data?

6. On which chromosome is the gene located? In what year was the gene first mapped to this chromosome?

7. In what year was the gene first identified?

(ii) Locating information on your chosen gene

In this portion of the exercise, you will locate specific information on your chosen gene. In order to locate this information, you must click on the hyperlinked "Links" located at the top right-hand corner of the record you selected above. Once you have clicked on "Links", a menu should appear. In this menu, select "Gene". Once you have selected "Gene", an annotated record that pertains to your gene will appear. Using this record, please provide answers to the following questions:

1. How many exons and introns are present in this gene?

To locate this information, please scroll down until you see a section entitled "Genomic regions, transcripts and products". An image should be shown. On top of the image, a sequence identifier for this portion of DNA should be shown and should be underlined. If you click on this name, a "Nucleotides Links" menu should appear. Select "GENBANK". The GenBank record for your gene should appear. Scroll down until you see "CDS", which stands for coding sequence. From the information located in this portion, you should be able to determine the number of exons and introns in the gene.

2. What is the length of the mature mRNA for this gene?

To locate this information, please scroll down until you see a section entitled "Genomic regions, transcripts and products". An image should be shown. On the left of the image, a sequence identifier for this portion of DNA should be shown in blue and should be underlined. If you click on this name, a "MRNA Links" menu should appear. Select "GENBANK". The GenBank record for the mRNA of your gene should appear. The length of the mature mRNA should appear at the top of the record.

3. What are the names of the genes adjacent to this gene on the chromosome?

To locate this information, please scroll down until you see a section entitled "Genomic context". An image should be shown. From the information located in this section, you should be able to determine which genes are adjacent to your chosen gene on the chromosome.

4. On which chromosome is the gene located?

HERE, YOU CAN DOUBLE-CHECK THE ANSWER YOU GOT EARLIER! To locate this information, please scroll down until you see a section entitled "Genomic context". An image should be shown. From the information located in this section, you should be able to determine on which chromosome the gene is located.

(iii) Locating information on the protein resulting from your chosen gene

In this portion of the exercise, you will locate specific information on the protein encoded by your chosen gene. In order to locate this information, you must click on the hyperlinked "Links" located at the top right-hand corner of the record you selected for part (A). Once you have clicked on "Links", a menu should appear. In this menu, select "Protein". Once you have selected "Protein", a list of records should appear. Select the protein that has been isolated from humans (*Homo sapiens*). Select this record using the hyperlinked sequence identifier.

Copy the sequence of your protein. It will be located at the very bottom of the record you selected above. Go to the following website: http://ca.expasy.org/tools/protparam.html. The program ProtParam is a tool that allows a user to calculate various physical and chemical parameters for a given protein.

Using ProtParam, please provide answers to the following questions:

1. How many amino acids are found in the wild-type form of the human protein encoded by this gene?

2. What is the predicted molecular weight of the protein?

(iv) Finding whether structural information is available for the protein

As we have briefly mentioned in class, a protein's function is intricately linked to its structure. The loss of a protein's three-dimensional structure usually results in a complete loss of function of the protein.

Protein structural information is very important to understanding the function of a protein. In many cases, the way in which a protein functions remains unknown until structural information is available. This is particularly true for enzymes, which are involved in the catalysis of various reactions. Due to the importance of structure in defining function, pharmaceutical companies invest many research dollars towards structure determination.

Protein structure determination can be done through both laboratory techniques and bioinformatics methods. Laboratory techniques include the use of X-ray crystallography and nuclear magnetic resonance (NMR). In X-ray crystallography, small crystals of the protein are grown and placed in the path of a beam of X-rays. Computer analyses of the resulting diffraction patterns reveal the three-dimensional structure of the protein. In NMR, a solution of protein is placed in a strong magnetic field. The interactions of the magnetic moments of the atoms of the protein and the magnetic field are monitored to deduce the three-dimensional structure of the protein.

The X-ray or NMR structures of proteins are stored in a publicly accessible database known as the RCSB Protein Data Bank (http://www.rcsb.org/pdb/home/home.do). This database contains

all of the three-dimensional coordinates for proteins, and allows a user to freely download the coordinates of a protein, and study and visualize the structure on a personal computer.

Thousands of protein structures are now available from the RCSB Protein Data Bank. The tremendous growth in the availability of protein structural information has helped fuel the popularity of bioinformatics methods. Bioinformatics methods are based on the use of known protein structures (from the protein data bank) to predict the structure of a similar protein. The last decade has seen the introduction of many software programs that can accurately predict the three-dimensional structure of a protein based on the structure of a similar protein.

To search for structural information on the protein encoded by your chosen gene, you will be using BLAST®. The website for BLAST describes this program as follows (http://www.ncbi.nlm.nih.gov/blast/blast_overview.shtml):

"BLAST® (Basic Local Alignment Search Tool) is a set of similarity search programs designed to explore all of the available sequence databases regardless of whether the query is protein or DNA (query is simply the sequence that you are providing to compare against the sequences in the database). The BLAST programs have been designed for speed, with a minimal sacrifice of sensitivity to distant sequence relationships. The scores assigned in a BLAST search have a well-defined statistical interpretation, making real matches easier to distinguish from random background hits. BLAST uses a heuristic algorithm that seeks local as opposed to global alignments. Therefore, it is able to detect relationships among sequences that share only isolated regions of similarity."

Essentially, BLAST is a program used to find sequences (nucleotide or amino acid) in a database that are similar to your query sequence. It also provides a score based on the degree of similarity between the sequences, allowing you a measure of the extent of similarity.

When is BLAST used? Let's look at an example. Let us assume that you have just cloned and sequenced a human gene. You want to see if this gene is similar to any other genes in the genome, or to genes in other species. Let us say you do a BLAST search and you find that your gene is very similar to other genes that encode protein kinases (proteins that phosphorylate other proteins). This would suggest that your gene also encodes a protein kinase. Therefore, your first experiment may be to test your protein for kinase activity.

There are different types of BLAST searches that vary primarily on two factors:

- Is the query sequence nucleotide or protein?
- Do you want to search databases that contain nucleotide sequences or ones that contain protein sequences?

We will only discuss three different types of BLAST searches (there are many more, but these are the main ones). First, click on the BLAST link at the top of the NCBI homepage. This should bring you to the following website: http://www.ncbi.nlm.nih.gov/blast/Blast.cgi.

1) **Nucleotide blast (blastn)** – This is when you have a nucleotide sequence (query) and you want to search a nucleotide database. To perform this type of search, click on the hyperlinked "nucleotide blast". The basic search just consists of typing in (or more likely, cutting and pasting) a nucleotide sequence in the upper box and pressing the BLAST button. We will not try this feature, but rather try the next type of BLAST search.

2) **Protein blast (blastp)** – This is when you have a protein sequence and want to search a protein database. To perform this type of search, click on the hyperlinked "protein blast". This page looks virtually identical to the blastn page. Again, a basic search is performed by typing in (or more likely, cutting and pasting) a sequence, but this time it is a protein amino acid sequence, and then pressing the BLAST button.

3) **blastx** – This is the third type of blast that we will briefly cover. This is when you have a nucleotide query, but you search a protein sequence database. The software automatically translates the DNA into an amino acid sequence, and then searches the database. The tricky part about this type of search is that the DNA is translated in six different reading frames. This is because either the top strand or the bottom strand of DNA could be the coding strand. Moreover, for each strand, the coding frame can start at three different positions. For example, the sequence agaattacctcagat could be "aga att acc tca gat", or "a gaa tta cct cag at" or "ag aat tac ctcaga t". Each of these would give a completely different amino acid sequence.

For this exercise, you will be asked to do a protein blast (blastp) to search the RCSB Protein Data Bank for structural information. Using the protein record you found in the last exercise, highlight and copy the sequence of the protein into the query window of the blastp site. Do not worry about selecting spaces and sequencing numbers – the search software will ignore these. Under the "Choose Search Set" tab, select "Protein Data Bank proteins (pdb)" under the Database drop-down menu. Press the "BLAST" button.

The first page to come up shows you some of the regions of the protein that are known to have functions in other proteins. The presence of these domains can give you a lot of information as to the function of the protein. **Make note of these putative conserved domains and include a copy of the figure in your poster.**

Although this page contains a lot of useful information, it is not the actual result page. Blast searches can take some time, depending, in part, on how many people are trying to use it at the same time. Remember that you are searching your query against billions of other sequences. This takes an enormous amount of computing power.

On the result page, near the middle of the page, is a colorful representation of your blast results. The thick red line shows the length of your query sequence. The thinner lines underneath are the best blast hits (proteins in the database with which it aligns well). If the hits are very good (very similar), the lines are shown in red. Hits that are not quite as good are shown in purple – and so on as the color key indicates. The length of each line shows where the similarity is found relative to the query sequence.

Select the entry with the highest score. Click on the hyperlinked pdb access code located to the left of the record. Does the record contain structural information for the entire protein of interest? If not, which portion of the protein is contained within the record?

Week 4: Preparing and presenting a research poster

With your laboratory partner, you will present a poster in front of your peers in your laboratory section. Your poster should include the following:

- background information you gathered on your disease of interest from the "Genes and Diseases" collection, as well as from your selected review articles;

- ***all information*** you gathered on your gene and protein using OMIM and BLAST, paying special attention to the inheritance pattern of your chosen disease, as well as the molecular basis for the disease (e.g., you are expected to explain to your peers how the available molecular data explains the Mendelian inheritance pattern of the disease); and
- a short description of the patent you located (including all information you were asked to locate from the patent).

Your poster should be made using a **single** poster board. The dimensions of the poster must not exceed 56 cm x 71 cm. Please remember that posters are VISUAL summaries and are meant to present data in an appealing fashion. Your poster should not consist only of writing. Please include as many figures and drawings as you feel are necessary. Also, be careful about font choice and font size. Your poster should be able to be read from at least 3 feet away. The use of color is always recommended to make your poster more attractive to read. Remember that your poster is meant to draw the attention of fellow researchers and engage them in a discussion.

During your poster presentation, you should clearly describe your chosen genetically-inheritable disease to your peers. Your presentation should be no shorter than 5 minutes, and no longer than 10 minutes. Remember to speak slowly and clearly. Your goal is to teach your peers the aspects of your chosen disease that you feel are most important for others to understand.

Your poster and poster presentation will be evaluated using the attached rubric (see APPENDIX B). Please note that the rubric does award marks for asking 2 questions during the entire poster presentation session. This means that, to be awarded these marks, you must ask 2 thoughtful questions during the session. The 2 questions must be asked on different presentations.

Materials

A computer having an Internet connection is required for weeks 1, 2 and 3 of the exercise per pair students. Each pair of students will require a 56 cm x 71 cm poster board for week 4.

Notes for the Instructor

One of the major challenges that we faced in implementing an inquiry-based exercise in a large class of over 500 students was to organize the exercise in a way that maximized the inquiry experience of each student without placing excessive demands on the limited time and resources of a small team of graduate teaching assistants, librarians and instructors. Several design elements of the exercise were specifically chosen to meet this significant challenge.

First, the exercise was integrated into four of the regularly scheduled, weekly laboratory sections of the introductory genetics course. The relatively small groups of students in laboratory sections (approximately 24 students in each of 21 laboratory sections) facilitated the interactive nature of the computer-based sessions during weeks two and three by providing opportunities for one-on-one interactions with teaching assistants and librarians, as well as peer-to-peer learning opportunities.

A second key design element was to encourage a student-driven, inquiry-based learning experience. The students enjoyed making the initial choice of their research topic from a broad list based on personal interests. In addition, it was particularly important to provide an open and flexible learning environment during the library and bioinformatics sessions to encourage student-initiated

inquiry-based learning. Sessions were geared towards explaining the technical aspects of bioinformatics tools and databases through the use of cystic fibrosis as an example of a human genetically-inheritable disease. Cystic fibrosis was specifically excluded from the list in APPENDIX A to allow for its use as an example during help sessions. The computer laboratories were equipped with an LCD projector, which allowed a librarian to easily demonstrate the use of the bioinformatics tools and databases to locate Mendelian and molecular genetics data on cystic fibrosis. Students were required to apply their basic understanding of classical Mendelian and molecular genetics to effectively use these tools to find the information necessary for their chosen disease. The combination of librarians and graduate teaching assistants leading small groups in a somewhat unstructured setting was particularly effective for guiding students through the technical and conceptual aspects of this inquiry-based discovery process.

Over the last two years, it has been our experience that the success of the inquiry-based discovery process depends on a few critical factors. First, bioinformatics and other non-bibliographic resources are complex and generally contain more information than students need in an introductory genetics course. During the help sessions, students are shown how to focus on the most useful aspects of the tools and databases, thereby greatly reducing frustration and encouraging later exploration of other aspects of these resources. Second, the exercise should try to emulate the way in which an expert would use these tools. Using these resources to ask "real" questions lends credibility to the exercise. Third, ensure that sufficient time is provided for hands-on practice with readily available assistance from the librarian, instructor, or teaching assistant. Questions often arise regarding searching and interpreting data that require expertise.

The third important design element essential for the success of the exercise was the use of a combined poster/oral presentation as a final evaluation tool. It was particularly effective to use the final laboratory session for student presentations and discussion. Most students viewed the final presentation of their results as a meaningful capstone learning experience. Through this part of the exercise, students had to organize and present their work in a way that demanded a deeper level of understanding than normally required in a typical exam or laboratory report.

One of the most difficult challenges facing this project was to devise a way to evaluate how students performed in the inquiry-based exercises. Since the combined poster/oral presentation was designed to be the culmination of the student-initiated inquiry-based learning process, the overall performance of the students in this exercise was evaluated by marking the quality of the poster/oral presentations for each pair of students. To deal with the large number of students typical of core undergraduate biology courses, we prepared a detailed marking rubric that provided specific guidance to the graduate teaching assistants regarding the grading of the final student presentations (APPENDIX B).

The marking rubric was carefully designed to emphasize the importance of creativity and inquiry, as opposed to a non-selective listing of information. Students were informed well in advance of their presentations that they would be marked for their creativity and the quality of their presentation, as well as for the scientific accuracy and completeness of information. As a result, students needed to master basic concepts and apply them in a meaningful way to prepare a successful presentation. To prepare their final presentation, each pair of students synthesized their abstract knowledge and applied it towards explaining a specific disease. This was the first opportunity for most of the students to present scientific information in a formal setting, and it proved to be a challenging but rewarding experience for most of them. As an added incentive, we informed the students at the beginning of the exercise that the students who prepared the best three

posters would be selected to participate in the annual departmental research symposium. The high quality of most of the presentations attests to the success of this method of evaluation, as well as the success of the project as a whole.

List of Websites

BLAST: http://blast.ncbi.nlm.nih.gov/Blast.cgi
Canadian Intellectual Property Office: http://patents1.ic.gc.ca/intro-e.html
ExPASy: http://ca.expasy.org/tools/protparam.html
Google Patents: http://www.google.com/patents
NCBI: http://www.ncbi.nlm.nih.gov
OMIM: http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim
RCSB Protein Data Bank: http://www.rcsb.org/pdb/home/home.do

Acknowledgements

Financial support is gratefully acknowledged from the University of Calgary Provost's Teaching and Learning Fund, the Department of Biological Sciences and the Faculty of Science. We also thank all of the undergraduate students at the University of Calgary who have taken part in this inquiry-based exercise and provided us with invaluable feedback. Their hard work and enthusiasm made this new exercise possible. We also thank Alexander Hynes for technical assistance.

Literature Cited

Bednarski, A. E., S. C. Elgin and H. B. Pakrasi. 2005. An inquiry into protein structure and genetic disease: introducing undergraduates to bioinformatics in a large introductory course. *Cell Biology Education* 4: 207-20.

Bowling, B. V., C. A. Huether and J. A. Wagner. 2007. Characterization of human genetics courses for nonbiology majors in U.S. colleges and universities. *CBE Life Sciences Education* 6: 224-32.

Boyle, J. A. 2004. Bioinformatics in undergraduate education: Practical examples. *Biochemistry and Molecular Biology Education* 32: 236-238.

Dymond, J. S., et al. 2009. Teaching synthetic biology, bioinformatics and engineering to undergraduates: the interdisciplinary Build-a-Genome course. *Genetics* 181: 13-21.

DeHann, R.L. 2005. The impending revolution in undergraduate science eduation. *Journal of Science Education and Technology* 14: 253-269.

Feig, A. L. and E. Jabri. 2002. Incorporation of bioinformatics exercises into the undergraduate biochemistry curriculum. *Biochemistry and Molecular Biology Education* 30: 224-231.

Haffie, T. L., Y. M. Reitmeier and D. B. Walden. 2000. Characterization of university-level introductory genetics courses in Canada. *Genome* 43: 152-9.

Holtzclaw, J. D., et al. 2006. Incorporating a new bioinformatics component into genetics at a historically black college: outcomes and lessons. *CBE Life Sciences Education* 5: 52-64.

Miskowski, J. A., D. R. Howard, M. L. Abler and S. K. Grunwald. 2007. Design and implementation of an interdepartmental bioinformatics program across life science curricula. *Biochemistry and Molecular Biology Education* 35: 9-15.

Sheley, S. M. and T. R. Mertens. 1990. A Survey of Introductory College Genetics Courses. *The Journal of Heredity* 81: 153-156.

Voet, J. G., et al. 2003. Recommended curriculum for a program in biochemistry and molecular biology. *Biochemistry and Molecular Biology Education* 31: 161-162.

About the Authors

Isabelle Barrette-Ng has been an Instructor at the University of Calgary since 2006, where she teaches large courses in genetics and biochemistry, primarily at the second-year undergraduate level. She graduated from Queen's University in 1998 with a BSc (Hons.) in Biochemistry, received a MSc in Bioinformatics/Biochemistry from the University of Montreal in 2001 and completed her PhD studies in Biochemistry at the University of Alberta in 2003. Her interests include the development and integration of inquiry-based learning approaches in large undergraduate courses. In 2008, Isabelle was awarded the Student's Union Award for Teaching Excellence for the Faculty of Science at the University of Calgary.

Don MacMillan has been the Liaison Librarian for Biological Sciences, Math, Physics and Astronomy at the University of Calgary Library since 2003. He provides program-integrated information literacy instruction and advanced reference and training services to students and faculty in those disciplines and carries out research on student learning, information literacy and the incorporation of tools and technology in information literacy instruction. Prior to his present position, Don was EMBA/MBA librarian at the Haskayne School of Business, University of Calgary and held several positions at the Calgary Public Library. Don received his Bachelor of Science and MLS degrees from Dalhousie University, Halifax, Nova Scotia.

Dave Hansen received his Ph.D. in Genetics from the University of Alberta in 1999 studying sex determination in *C. elegans*. He then was a post-doctoral fellow in the Department of Genetics at the Washington University School of Medicine. During his post-doctoral studies, he studied the genetic control of proliferation in the *C. elegans* germ line. He joined the faculty of the University of Calgary in the Department of Biological Sciences in 2004, where he continues his research on *C. elegans* germline proliferation. He teaches developmental biology and genetics to undergraduate students, including Mendelian genetics in the second-year Introduction to Genetics course.

Appendix A: List of human genetically-inheritable diseases that can be used for the inquiry-based exercise

Adrenoleukodystrophy
Alzheimer disease
Amyotrophic lateral sclerosis
Anemia, sickle cell
Alpha-1-antitrypsin deficiency
Angelman syndrome
Achondroplasia
Alport syndrome
Adrenal hyperplasia, congenital
Autoimmune polyglandular syndrome
Ataxia telangiectasia
Atherosclerosis
Breast and ovarian cancer
Best disease
Burkitt lymphoma
Colon cancer
Crohn's disease
Cockayne syndrome
Charcot–Marie–Tooth syndrome
Deafness
Diabetes, type 1
DiGeorge syndrome
Duchenne muscular dystrophy
Diastrophic dysplasia
Ellis-van Creveld syndrome
Epilepsy
Essential tremor
Familial Mediterranean fever
Fragile X syndrome
Friedreich's ataxia
Fibrodysplasia ossificans progressiva
Gaucher disease
Glaucoma
Gyrate atrophy of the choroid and retina
Glucose galactose malabsorption
Harvey Ras oncogene
Hemophilia A
Huntington disease
Hereditary hemochromatosis
Leukemia, chronic myeloid
Lung carcinoma, small cell
Lesch-Nyhan syndrome

Long QT syndrome
Malignant melanoma
Marfan syndrome
Myotonic dystrophy
Male pattern baldness
Menkes syndrome
Maple syrup urine disease
Multiple endocrine neoplasia
Obesity
Immunodeficiency with hyper-IgM
Neurofibromatosis
Niemann–Pick disease
Narcolepsy
The p53 tumor suppressor protein
Paroxysmal nocturnal hemoglobinuria
Pancreatic cancer
Polycystic kidney disease
Porphyria
Prader-Willi syndrome
Prostate cancer
Pendred syndrome
Parkinson disease
Phenylketonuria
Retinoblastoma
Refsum disease
Rett syndrome
Severe combined immunodeficiency
SRY: Sex determination
Spinal muscular atrophy
Spinocerebellar ataxia
Thalassemia
Tangier disease
Tay-Sachs disease
Tuberous sclerosis
Von Hippel-Lindau syndrome
Wilson's disease
Williams syndrome
Waardenburg syndrome
Werner syndrome
Zellweger syndrome

APPENDIX B: MARKING RUBRIC FOR POSTERS AND POSTER PRESENTATIONS

Name:_____
Topic:_____

Content (5 marks):

Give 1 mark if that aspect is well done, 0.5 marks for evidence of effort, and 0 for no effort.

_____ clearly state chosen genetically-inheritable disease
_____ information presented is clear and coherent, demonstrating good understanding and knowledge
_____ appropriate amount of background information is presented
_____ inheritance pattern and molecular basis for disease are clearly explained
_____ references are properly cited on poster

Presentation (7 marks):

Give 1 mark if that aspect is well done, 0.5 marks for evidence of effort, and 0 for no effort.

_____ demonstrate good speaking skills (tone, volume, pace, avoiding "ums", "you know", etc.)
_____ logical and confident delivery of material
_____ presentation is well-organized
_____ presenters make eye contact, engage the audience and are euthusiastic
_____ poster is clear and can be easily read
_____ poster is appealing, interesting and inviting
_____ equal participation of group members

Answering questions (2 marks):

2 marks will be given if all questions are answered thoroughly
1 mark will be given if the questions are only partially answered
0 marks will be given if the questions are not answered satisfactorily

Asking questions to others (2 marks):

2 marks will be given if 2 thoughtful questions are asked
1 mark will be given if only 1 thoughtful question is asked or if the questions are not thoughtful
0 marks will be given if no questions are asked

Deduct 1 mark for every minute that the talk goes over 10 minutes or for every minute that is missing to reach a minimum of 5 minutes.

Poster and presentation mark = _____/16

Comments: