Integrating Bioinformatic Instruction Into Undergraduate Biology Laboratory Curriculum

Diane C. Rein[1], Jennifer Sharkey[2] and Jane Kinkus[3]

Purdue University Libraries
504 W. State St.
West Lafayette, IN 47907

[1]Life Sciences Library
drein@purdue.edu

[2]John W. Hicks Undergraduate Library
sharkeyj@purdue.edu

[3]Mathematical Sciences Library
jkinkus@purdue.edu

Abstract: This computer-based laboratory workshop deals with teaching the concepts of bioinformatic research discovery at the undergraduate level in a practical, hands-on, active learning experience that mimics primary bioinformatic bench research. The intent is to integrate bioinformatic literacy into existing laboratory exercises. The computer exercises are divided into four modules of increasing complexity that cover finding information at NCBI, using PubMed's interface as a paradigm for searching and linking to information at NCBI, learning to globally search and analyze bioinformatic information at NCBI's Entrez and finally, learning to effectively hyperlink search and analyze gene structure and function information through Entrez Gene.

Introduction

In the past decade, the life sciences research discipline has been transformed by the exponential growth of web-based publicly available genetic information databases as a direct result of the Human Genome Project. Coupled with advances in computational methods and genomic technologies, these new, often confusing multitude of Internet-dispersed "biological databases" (i.e., GenBank, OMIM, Ensembl, FlyBase, KEGG), now contain a rich array of extremely large datasets which, in turn, have spawned secondary and tertiary databases mounted on the Internet by both organizations and individual researchers. These biological databases, blending biology with computer technology (bioinformatics) represent a new kind of scientific literature quite distinct in access and purpose, from the traditional scholarly publications indexed in literature databases such as PubMed. They hold immense amounts of complex biological data elements ranging from raw data to curated literature in a highly interlinked electronic environment. Unlike the print literature, students of biology now have to deal with very complex and large datasets that often are presented with little context or meaning.

What is Bioinformatics?

Bioinformatics is a research process where the computer is the primary research tool used to access Internet-based data and information pertaining to biological molecules, both macromolecular (DNA, RNA and protein) and micromolecular (metabolites, amino acids, drugs, ions, etc.). Researchers typically access these databases to either develop research protocols, hypotheses and strategies at the

laboratory bench, or they enter bioinformatic resources to check on and validate results from "wet-bench" research. This can include computationally finding or confirming DNA sequence data, determining putative functions and interactions of DNA, RNA or proteins through sequence analyses, associating a disease or pathological processes to a sequence isolated in the laboratory or to the use of tools such as electronic PCR, BLAST sequence searching, computationally determining how to knock out gene function, locating putative gene regulatory sequences, or computer-based research studies to determine if a particular mutated sequence will alter the function of a protein before attempting to create the mutation and working with it *in vivo*. Bioinformatics research has become so embedded into the everyday research processes of life sciences research today, with investigators flipping back and forth between computer-based research and "wet-bench" research, that bioinformatics IS life sciences research today.

Bioinformatics as a discipline

Bioinformatics is also those professionals who create the software to perform bioinformatic research, as well as the algorithms, mathematical expressions and statistical packages needed to process and retrieve bioinformatic queries. Understanding mathematics and having some knowledge of simple computer programming skills, such as PERL scripting, or in database management and organization, are also areas that biological laboratory educators should be considering as a part of the undergraduate laboratory experience. This is as important in the non-major biology curriculum as in the biology major curriculum: Those students who may be the next bioinformatic computer/software engineers may not be biology majors at all.

Bioinformatics is a paradigm

Bioinformatics represents a shift in scientific investigation from the journal article (information in context) to raw datasets themselves. Curated datasets found on the Internet, represent a new type of scientific publication and scholarly communication. Educators in biology have to reach out from the textbook to these new kinds of database "publications". Sequence searching (BLASTing) is actually a form of data-mining and pattern recognition. Unlike analyzing the results from a literature (text-based) search, analysis of many kinds of bioinformatic search results requires some savvy in mathematical thinking and statistical analyses.

Biological Science as Information Science: Implications for the Undergraduate Biology Laboratory Experience

Bioinformatics, as the "New Science," as "digital biology," is transforming biology into an information science. Research now carries with it a strong search component, pressuring both students and their instructors to learn fundamental informatics in order to perform *in silico* research that supports the traditional wet-bench research process. Any existing wet-bench undergraduate laboratory course and/or module could conceivably inject a bioinformatic component, mimicking how research is carried out today in the sciences. To this end, the published literature contains a variety of excellent tutorials in bioinformatics (Almeida, *et al*., 2003; Bednarski *et al*., 2005; Honts, 2003; Krawetz, 2000; Kumar, 2005; Mulniz, 2003; Ranganathan, 2005; Rice *et al.*; Weaver *et al.;* Wefer, 2003). Often these are stand-alone modules, separated from existing biological laboratory exercises instead of embedded within them as a necessary part of the overall research experience. Moreover, from an entry level undergraduate student's point of view, these can be advanced exercises and quite foreign in design and concept.

In addition to learning about biological processes and organisms and performing wet-bench research, biology students today have to develop information science skills. They have to learn to think informatically. How are the data organized? How does one set of data relate to others? How does one migrate between datasets, search within databases, deal with a myriad of differing interfaces and then interrelate these data to the traditional published literature? How does one analyze the results, discerning patterns, and then place them into the context of the research question posed?

Integrating Bioinformatic Instructional Modules Into Existing Undergraduate Laboratory Curricula

Most existing undergraduate laboratory courses have a series of discrete, self-contained experiments to illustrate specific tools and research methodologies in biological science. Thus there could be a module on DNA sequencing, polyacrylamide gel electrophoresis, measuring protein levels, the polymerase chain reaction, various enzymatic assays or studying Mendelian inheritance patterns with *Drosophila*. Bioinformatics research, however, is systemic to the life sciences and not restrained to a particular subject area or technique. In each and every one of the examples above, a bioinformatic module could be created that took students to the Internet to perform research through searching the appropriate databases and using computerized tools.

It is the premise of this workshop, that bioinformatics is most effectively taught by giving students the experience of working with it as an integrated component of existing laboratory modules, most accurately mimicking the true primary research experience. Alternatively, bioinformatics could be taught as an additional, complementary laboratory for the entire academic year, or instructors could create discrete stand-alone modules that are complementary to existing laboratory curricula, but taught separately and simultaneously. With any of these approaches, each succeeding year of the undergraduate biology major's experience can be elaborated upon, added to, and extended upon, until the senior year when bioinformatic instruction and skills would reach a high degree of sophistication and level of expertise by the students. Because search is research in bioinformatics, it is quite conceivable to build a 4-year bioinformatic laboratory instructional program that could aim for students to be performing primary research that results in independent research publications in their last year as undergraduates. Regardless of the format chosen, it might be best to view bioinformatics laboratory undergraduate instruction holistically—as a four-year experience and an integrated component of a department's entire undergraduate biological curriculum—rather than as a discrete set of exercises spanning several weeks in undergraduate laboratories, or confined to one or two laboratory courses.

It is the purpose of this tutorial to give a fundamental bioinformatic experience to undergraduates enrolled in biology undergraduate laboratory courses prior to moving on to more sophisticated bioinformatic techniques. Often overlooked is that these same problems and learning needs arise for faculty teaching bioinformatics. A gap exists between those who create and mount bioinformatic data resources to the Internet and those who must teach them who have to gain a higher level of understanding of the resources than their students. This tutorial is centered to those bioinformatic resources at the National Center for Biotechnology Information (NCBI, http://www.ncbi.nih.gov/). Sooner or later, students learning bioinformatics have to learn and deal directly with at least NCBI, which contains a large set of interrelated databases. Information regarding a given protein, gene or disease could easily be spread across a dozen different databases. NCBI, therefore, is an excellent paradigm for developing basic information seeking skill and understanding what kinds of bioinformatic databases exist and their function, whether they be at NCBI or scattered throughout the Internet.

Student Outline

Module 1: Learning Basic NCBI Web Site Navigational Skills

The National Center for Biotechnology Information (NCBI) at http://www.ncbi.nlm.nih.gov, housed within the National Library of Medicine at NIH, is the United States national resource for molecular biology (bioinformatic) information. NCBI collects data and information related to DNA, RNA and proteins, organizes them into over 70 different databases and tools and makes this information publicly available. Before using NCBI's resources, you must first learn what resources are available at NCBI, how to locate them, and what associated help documents exist for each one. Each help document covers different sources in different ways. While they often overlap in content, each also contains unique information. Collectively, they cover all the resources available at NCBI.

Below are the six major places to locate NCBI Help documents:**

1. NCBI Web Site Search. In the upper left hand corner of the NCBI home web page is a search box with a pull-down menu. Pull-down the menu, select "NCBI Web Site" and then enter a search term in the text box to the right. This search engine searches all the text on any of NCBI's web pages.
2. Below the search box, in the left-hand "blue navigational side bar" at NCBI is the "SITE MAP". Click on the text and it will take you to a visual map of the NCBI website.
3. Below the SITE MAP is the "Alphabetical List". Clicking on this text will take you to a selected list of *database* links. Clicking on each link will take you directly to each individual database's home page where more help information will be available, typically in the left-hand blue navigational bar.
4. Below the Alphabetical List is the "Resource Guide". Clicking on this text takes you to brief description of each resource (databases and other resources types) and then to the home page of the resource itself.
5. Towards the bottom of the left-hand blue bar on NCBI's home web page, is the "Education" section. Clicking on this text will take you to tutorials, many help documents and other educational material.
6. Using one or all of the above five resources, you will eventually find your way to the home page of each individual database, tool or resource at NCBI. Often (but not necessarily always) Help documents specific to that database or resource will also be located in the left-hand blue navigation bar at each database.

**You can rapidly search for information in any given web page by using the "find in this page" command within your web browser, usually located in the Edit Menu of the browser.

Using any combination of the resources above, and only these resources, *answer the following questions:*

1. Locate the "NCBI Help Manual" and give the URL for it. Briefly describe what it is and what information it covers.
2. Locate the "NCBI Primer". Provide the URL and give a brief description of what it contains.
3. What is "Genes and Disease"?
4. What is the NCBI Data Model? Using the Data Model, list the databases to which the Gene database is hyperlinked.
5. What is the Entrez set of databases?
   - What is the URL for the home page of Entrez?

- Give a brief description of Entrez and where you obtained this information at NCBI.
- There is more than one way to migrate to Entrez at NCBI. Name at least three of them.
- Select 8 different resources within Entrez that interest you and give a 3-5 sentence description.
- Locate the Entrez Tutorial and provide its URL.
- Locate the Entrez Help and provide its URL.

6. Compare and contrast the RefSeq and GenBank resources at NCBI. What are the URLs for each?
7. What is BLAST?
   - List the different kinds of BLAST resources at NCBI.
   - Provide the URLs to their home pages.
   - How do they differ from one another?
8. What resource(s) does NCBI have concerning RNAi? Give the URLs for any you find.
9. What are PubMed, OMIM and OMIA? Compare and contrast them.
   - Provide the URL for each.
   - How are they different from the other databases at NCBI?
10. What are Cn3D and MapViewer at NCBI?
    - Provide the URL for each
    - How do they differ from the other databases at NCBI?

Module 2: NCBI PubMed—Searching for Gene Information in the Literature

PubMed is the public research literature database and is a component of NCBI's Entrez databases. Although it originally was created to collect articles in the biomedical research literature, in the last five years it has expanded its coverage to include most of the sciences. One can search the database for information on genes, mutations, pharmacology, nutrition, diseases, plants, animals, psychiatry, climate change, astronomy, biomedical ethics, alternative medicine and much more. PubMed existed prior to the Human Genome Project and the massive sequencing of DNA and genomes. When it came time to mount the sequencing data to the Internet, the PubMed interface was chosen for the NIH bioinformatic databases at NCBI. As a result, the Entrez databases use the PubMed interface and functional features as a foundation. If you learn to use PubMed's features effectively, you can easily transfer your new knowledge to searching all of the NCBI bioinformatic databases in general.

Details Tab Exercise

Go to the home page of PubMed at NCBI, by going to NCBI's home page and clicking on PubMed in the upper left hand corner.

1. In the search box, type in "bad breath", but do not use the quotation marks. Click on "Go" or hit the Enter key on your keyboard. The search will automatically run.
   a. Click on the Details Tab. The Details shows you how PubMed ran the search. If it found a matching term that it recognized in its vocabulary database called MeSH (Medical Subject Headings), you will see a search term in quotation marks and the phrase: [MeSH Terms]. In this search it found that "halitosis" is the medical term for your search term "bad breath". You now know for future reference that when you run searches for bad breath in any database you can perform this search, "halitosis OR bad breath", instead of using the one search term "bad breath" only.

2. You can edit the search here in the Details box. To do this, remove all the search terms except: "halitosis"[MeSH Terms] in Details search box. Click on the "Search" button below the Details search box and the search will run again.
   a. You will now find that PubMed automatically added "halitosis"[MeSH Terms] to the Search box and that you have less hits in the results. The reason for this is because PubMed automatically converted your original search term of "bad breath" to the MeSH term AND simultaneously searched all the text in the Title and Abstract fields of all its 15 million articles in its database.
   b. Show the first 500 results on one page by using the "show" pull-down menu above the first result.
      i. Go to the 500th result. What is the reference?
      ii. Scroll to the top of the search results.
      iii. Pull down the "send to" menu.
      iv. What options are available? List them and briefly describe what they do by investigating each one on your own.
3. Clear the search box if needed and then type: maple tree. Run the search.
   a. Click on the Details Tab. PubMed automatically matches this to term "acer". What is that?
   b. Edit the search string in the Details Tab search box to contain only the MeSH terms like this: acer [MeSH Terms] AND "trees"[MeSH Terms].
   c. Click on the Search button BELOW the Details Tab text box. What happens? Editing your search string in the Details Tab is one way to force PubMed to focus your search to more relevant results.
   d. Sort the results by author, by using the pull-down menu for sorting.
4. Clear the search box and then run the searches listed below these instructions in PubMed. For each search, use the Details Tab and its Search button to determine the actual scientific or biomedical name for the term searched; then edit the Details box to force PubMed to only use any MeSH terms that it had originally found. Record the number of hits returned for both the edited and unedited searches. Which search (edited or unedited) works best in each case and why?
   a. Shingles
   b. Hypertension
   c. Cancer
   d. Breast Cancer
   e. Causes
   f. Fruit Fly
   g. Wisdom teeth

History Tab Exercise

The History Tab saves any of the searches you have run for up to eight hours. It only recognizes your searches from one computer at a time. If you are running searches at PubMed in your laboratory class computer and return home to use your personal computer, you will not be able to see any of the searches you ran elsewhere, including in your class. The History Tab also permits you to re-run searches and combine any searches already run. In this way, you can very easily build a complicated search string. You should never use the Back button in your browser to return to previous searches. PubMed only keeps your searches in the browser's history temporarily. The History Tab is the only way you can insure access to all of your previous searches.

1. Click on the History Tab.

a. Click on the "Clear History" button at the bottom of the page to remove all of your previous searches.
b. In the Search box, type in "hypertension" and run the search. How many results were returned and what is the medical name for "hypertension"?
c. Clear the search box.
d. Type in a new search term: nosebleed. Run the search. How many articles are there and what is the medical term for "nosebleed"?
e. Click on the History Tab. You will see both of your searches collected here. If you click on the extreme right hand link under "results", PubMed will run that search again. Note that the last search you ran, "nosebleed" is still in the search box.

2. In the History Tab, locate your search on "hypertension". LEFT-click with your mouse on the number next to it.
   a. From the pop-up menu, select AND. PubMed will now add the two searches together. The AND command instructs PubMed to find articles that contain both of your search terms, hypertension and nosebleed, in the same article. Note that both of the search terms are now in the search box. Run the search.
      i. How many results did you get?
      ii. Reading only the titles of each hit, what can you say about nosebleeds and hypertension? Write a short paragraph, citing the appropriate references within the text and creating a bibliography at the end of the paragraph of the articles you cited.

Limits Tab Exercise

The Limits Tab is a set of commonly used limits in PubMed. It is NOT the complete set. NCBI offers the Limits Tab as a quick access tool. The full range of the limits available in PubMed is actually in the Preview/Index Tab which we will use later. However, the Limits Tab contains a few highly useful search restrictions not found elsewhere within PubMed

1. Clear the History with the "Clear History" Button.
2. Click on the Limits Tab and look over the possible choices. Remember, these are limits that are commonly used in PubMed over the years.
   a. You are writing an article on the ethical issues of HIV and AIDS in adult women (over the age of 19). You only are interested in review articles on the topic in the last five years.
      i. Type HIV into the search box. Run the search and check the Details Tab to see how PubMed mapped your search request to its vocabulary. You will note that PubMed did not search for the concept of AIDS. Why not?
      ii. Clear the search box.
      iii. Run a new search for AIDS. Did PubMed run it correctly for you? If not, edit the search in the Details Tab until you are satisfied.
      iv. Use the History tab to combine your AIDS and HIV searches.
      v. Click on the Limits Tab. Locate and check the boxes for articles in the last five years, for women, and for ethics. Run the search. How many results did you obtain?
      vi. In the results, note that there is now a green check mark in the Limits tab. The limits you have applied will always be applied to any new searches until you uncheck this.
      vii. In the results page, note that there are two tabs: One reads "ALL" and one reads "Reviews". Click on the Review Tab and PubMed will now display the titles and abstracts of the review articles on this topic in the last five years.

Preview/Index Tab Exercise

The Preview/Index tab is the place where PubMed holds ALL of its available limit options.

1. Clear the History.
2. Clear the search box if necessary.
3. Make sure the Limits Tab is unchecked.
4. Without entering a search term, click on the Preview/Index tab
5. Scroll down if necessary until you locate a pull-down menu that says "All Fields"
6. Pull-down the menu and look over the various options available.
7. Go back to the top of the web page and type in "vioxx". Do NOT run the search yet.
   a. Click on the Preview/Index tab if you have not done so already.
   b. Scroll down the page and select "Publication Type".
   c. Highlight "Clinical trial, phase iv" in the scrolling text box that appears.
   d. Click on the "Index" button to the right of the Publication Type menu option.
   e. Click on the AND button under the "Publication Type" box.
   f. If you now go back to the top of the page, you will see that "type iv clinical trial" have been added to your search for "vioxx".
   g. At the top of the page, click on Go to run the search.
   h. Are there any phase iv clinical trials going on about Vioxx?
   i. Sometimes, running your search in PubMed gives you very few results. However, PubMed has an automatic "find related" articles link that searches for similar articles. Next to the article's abstract, click on the "See All Related Articles". How many more articles did PubMed find on Vioxx and clinical trials for you?
8. Your class has been given assigned topics in new methods in biology. You have to write a report on "charge coupled devices" and you have no idea what it is. You go to PubMed to try to find some research articles on it. The exercise below demonstrates how you can adjust your search to specific parts of a research article (*i.e.*, title, abstract) to focus the search to yield more relevant results for you.
   a. Before beginning, clear the History and uncheck the Limits Tab if necessary.
   b. Type in "charge coupled devices" (without the quotes) in the search box at the top of the page and run the search.
   c. The results do not look promising. The titles contain many topics. Reading the available abstracts is not much of a help. A few look like they have something to do with "charge coupled devices", but they are scattered through the 1100+ results. Click on the Details Tab. Did PubMed run the search the way you needed it done? Explain
   d. When using the Preview/Index Tab *to run a search*, you must first clear the search box at the top of the page and leave it blank. Do that now.
   e. Click on the Preview/Index Tab.
   f. From the "All Fields" pull-down menu, select "Title/Abstract".
   g. In the text box next to "Title/Abstract", type "charge coupled devices" (without the quotation marks).
   h. A new scrolling text box will open with all the terms that match "charge coupled devices". Select "charge coupled devices (1009)". The number means that there are 1009 possible articles to which PubMed has automatically mapped to charge coupled devices in its database.
   i. Click on the AND button below "Title/Abstract".

j. Go to the top of the page and you will see that PubMed created the search for you, but restricted it only to the use of the search term in either the Title or the Abstract of the research papers.
k. Run the search. Compare the number of hits with your first search. Is it more or less; better or worse? Why?
l. Clear the search box and repeat the above, but select "Title" from the "All Fields" pull-down menu. Compare the number of hits with your other two searches. Which of the three searches returns the most relevant results for you? Why would changing where PubMed searches within a journal article change the success of the search?

9. Sometimes you do not know how to run a search in PubMed or if PubMed would have the information you want. For example, you want to know if Vincent Van Gogh had ever been diagnosed with some kind of disease that may have affected his vision and give rise to his unique painting style. Because this search concerns a disease, it is very possible that this information might be collected in PubMed. Here is how to find out if PubMed has this information and if it does, to recover the articles. The trick is to use the Preview/Index tab.
   a. As for the above examples, clear the History and make sure the Limits Tab is not checked.
   b. Without running a search at the top of the page, click on the Preview/Index Tab.
   c. Scroll down the page and leave the Preview/Index pull-down menu as "All Fields, because you are not really sure where PubMed may hold this information in its database.
      i. Type into the search box here: van gogh.
      ii. Click on the Index tab to the right of the search box.
      iii. Scroll down the list looking for choices that match your search. List the various ones that seem to match.
      iv. Select one. Click on AND. It becomes your search at the top of the web page. Run the search and look over the results by examining the titles. Using the Display pull-down menu, select "citation" to read the abstracts, if available.
      v. Run the search for each term from the list you created above the search box at the top of the page, return to the clearing the search box at the top of the page, scrolling down to bottom of the page and selecting a new term by repeating steps i. through iv.
         1. Remember to use the "find all related articles" if needed.
         2. Did Vincent Van Gogh have a disease? What disease or diseases could he have had and how could it possibly have affected his painting style? Write a short paragraph or two on this topic, citing the articles you use and created a bibliography of the articles you cited.

Saving Searches or selected records in PubMed

The MyNCBI function of NCBI permits you to save searches in most of NCBI's bioinformatics databases, including PubMed. You can create multiple accounts with different user names and passwords and you can choose to have the searches run automatically on a regular schedule and emailed to you. No personal information is collected unless you want to have searches delivered to your email account, for which you must provide your email address. MyNCBI will save any search at any database within NCBI in one account. You can access all your saved searches from any of the NCBI databases.
To save your search string:

1. Clear the History and make sure the Limits Tab is unchecked.
2. In the search box, type Morgellons Disease and run the search.
3. A "save search" hyperlink appears to the right of the search box at the top of the page. Click on this.

4. If you already have a MyNCBI account and are not already signed in, you will be asked to log in. If you do not have a MyNCBI account, you will need to register. Create a User Name and Password and follow the instructions.
5. A pop-up box will appear, giving a default name to the search. Change it if you want, determine if you want to receive email updates or not, then click "OK".
6. The search is saved. You re-run the search anytime, by accessing your MyNCBI and clicking on the Searches Tab within it. The MyNCBI link always appears in the extreme upper right-hand corner at those web pages at MyNCBI that have the function available.
7. Re-run the search from within MyNCBI. What is Morgellons Disease and what is the controversy surrounding it? Write a short summary of your findings from reading the titles and any available abstracts to the articles. Cite the references used in the appropriate place within your article and create a bibliography of all references you used.

To save individual records from within a search:

1. Return to the Morgellons Disease search through the History Tab and re-run it.
2. Check the boxes to the left of the article titles for the first four articles.
3. From the "send-to" pull-down menu immediately above the search results, select "clipboard".
4. Click on the Clipboard Tab. You will see that your marked records have been imported into the Clipboard, which will save them for up to eight hours.
5. Pull-down the "send-to" menu, but this time select "MyNCBI Collections".
6. A pop-up box will again appear. You can create a "new collection" or append to one you have already created. Clicking "OK" will add the selected records to MyNCBI. You can access them by searching on the "Collections Tab" from within MyNCBI.

Module 3: NCBI Entrez—Searching for Bioinformatic Data

NCBI Entrez is a suite of 33 bioinformatic databases at NCBI that are hyperlinked together. The Entrez databases include PubMed which you have learned to use in the previous exercises. OMIM (Online Mendelian Inheritance in Man), OMIA (Online Mendelian Inheritance in Animals) and PubMed Central are also literature databases simultaneously containing bioinformatic data. Other databases (Nucleotides, Protein, SNP, etc.) are primarily data databases. In addition there are several other database and resources within Entrez. You can search across all the databases simultaneously at Entrez. Entrez will then attempt to map your search as best as possible to each of its databases and return the results for each. Unlike PubMed, however, the Entrez search engine does not automatically map your search term to a standard vocabulary. In order to search Entrez effectively, you have to learn to search a different way.

1. Go to the Entrez Home Page. You can do this several ways including from the Alphabetical List at NCBI's home page or by clicking on the "All Databases" text link in the horizontal dark blue bar that runs across the top of NCBI's Home Page.
2. At the next web page, you will see a table listing all the databases within Entrez. Before using Entrez, you need to learn what these resources are. If you click on each database name, you will be taken to the home page of each. If you enter a search string in the text box at the top of the web page, you will search across all the databases simultaneously. For this exercise, you will not need to enter any searches. Instead, using a combination of the Help documents you already located in Module 1, the Help document link at the Entrez page here, the short descriptions provided within Entrez for

each database or running searches in PubMed to learn about each database, answer the following questions. (You may have to click through to each individual database and investigate additional help documents):

a. If you want to search Entrez for DNA sequences, which database(s) would contain this information?
b. Which databases contain protein sequence information?
c. What is a SNP and how are they used in research?
d. What Entrez database will tell you what "Oryza sativa" is? How many Nucleotide records exist at NCBI for it?
e. What Entrez database would have information on Von Willbrand Disease in pigs?
f. Once you locate the Entrez database with this information, use this resource to locate more information for Von Willbrand Disease at NCBI. Write a short paragraph describing what it is and what causes it in animals. By sure to cite your sources in the text of the paragraph and create a bibliography of the references cited.
g. What Entrez resource(s) will information about the genes known (i.e. gene structure and function, not just the sequences or information only on the disease itself) to cause human muscular dystrophy?
h. What Entrez database(s) will have information on the chemical structure and biological activity of the anti-arthritic drug, Vioxx?
i. If you wanted to know what the three dimensional structure of a DNA polymerase looked like, what database at Entrez would have this information?
j. What are HomoloGene and UniGene? Compare and contrast these two databases. What information do they give about genes and proteins?
k. A common method to determine gene function today is to see what happens when that gene cannot function in the live animal. To do this, researchers "knock out genes" in a variety of ways which requires specific kinds of DNA and RNA sequence information to do so. What database at Entrez contains information related to this method?

3. Now that you've learned a little bit about what databases exist at Entrez, it is time to learn to search globally across Entrez.
   a. Go the Home Page of Entrez (i.e., the "All Databases" link) and in the search box at the top of the page, enter the search term you have already used in the PubMed exercise, Morgellons Disease". Click on the "Go" button to the right of the search box to run the search.
      i. Entrez returns the results for each database in a white or grey box. What do the white and grey boxes mean?
      ii. RIGHT-click on the results of the PubMed search. The search will now run in PubMed in a new tab in your browser. How do the results compare to when you ran this same search directly in PubMed in Module 2. Using the Details Tab is the search run in the same or differently than when run in PubMed directly? Did the search run properly?
      iii. Return to your original search results in Entrez, by clicking on the browser tab that holds the results. Now RIGHT-click on the Nucleotide results. At the next web page, click on the "Core Nucleotide" link. Now click on the Details Tab. How did this search run in Nucleotide? Did it run correctly? Why or why not?
      iv. Do the exercise in step iii above for ALL the search results in Entrez. Make a table of how each search ran in each of the Entrez databases and whether it ran the way it was supposed to run. What is your overall conclusion about how effective the global search

can be in Entrez with this particular search? After doing this exercise, explain in more detail the significance of the white and grey box search results for this particular search.

4. You will have now found out that Entrez sometimes cannot run a search well in many of its databases, in others it runs only part of your search string and in some the search runs properly. This is because, unlike PubMed, Entrez cannot map your search term automatically, and each database within Entrez has different Limits and Preview/Index Tab functions. You need to make use of the Details Tab on a regular basis when using the global search in Entrez to make sure you are obtaining the results you need. If not, you have to learn how to create a different search strategy, and modify it if necessary.
   a. For this exercise, we are interested in knowing about breast cancer in humans: What genes are involved, what is known about their proteins, something about the disease itself and anything else at NCBI that might help us know more about this cancer. Go to the Entrez home page and enter the following search string: breast cancer AND human. The "AND" is a Boolean Operator that mathematically combines the two search terms. Search results must contain both search terms someplace in the records. "AND" must be capitalized. When capitalized, AND becomes a software command, not a text word.
   b. When the results are returned at Entrez, click on the Nucleotide database link to get the results for this resource. At the next web page, click on the Core Nucleotides link.
      i. How many results do you have?
      ii. Save the results to your MyNCBI, giving it a unique name. You can use MyNCBI later to compare your saved searches for this exercise if needed.
      iii. Make a table of the name of the Entrez database and number of hits for each one.
      iv. Click on the Details Tab. The search ran this way: "Breast Cancer"[Journal] AND ("Homo sapiens"[Organism] OR human [All Fields]). Nucleotide ran the search mapping "breast cancer" to the name of a journal, not a disease. "Human" was mapped to the organismal field, which is correct. The square brackets are the limits set by the search. In PubMed this is done automatically for you. In Entrez, and almost all of its individual databases, you usually have to manually place or adjust the limits in yourself by typing in the correct field in square brackets after the search term. To the table you have created, make column for how the search ran in the Details Tab and how it should have run if in error.
      v. Click on the Entrez results for the Protein, Gene and OMIM databases and for each look at the Details Tab, and add the information to your data table about how this search ran in each of these databases as you have already done for the Nucleotide database.
   c. Let's modify this search. First, however, we need to know what the fields are in Entrez that can be placed in square brackets so we can choose them properly. Unfortunately, NCBI does not offer this information for ALL of the Entrez databases. Only for the most common data databases searched. It is available in the Entrez Help document, under the chapter titled "Using the Indexes", and the links to Table 1, 2 and 3. Open up these links, print them out if necessary and look them over before modifying the Entrez search string at its home page to: breast cancer[text] AND human[organism] instead of simply "breast cancer" and human.
      i. How is Entrez going to run this search now, compared to the one just run above?
      ii. Save your search results for this second search, with a different name from the one above, in your MyNCBI to work with later if necessary.

iii. In the table you created above, write down the number of hits you receive with this search and for what database in Entrez as you have done above for the first search.
iv. Click on the Details Tab for the Nucleotide, Protein, Gene, and OMIM databases and compare how the search ran in each database, by adding this information to data table you started for the "breast cancer" and human search.
v. In the Nucleotide database, locate WHERE in the first several records recovered, you find the search term "breast cancer" and indicate that in your data table. (You can do this easily if you used the "find" or "find in this page" command from your browser's Edit Menu). Is the search running finding your search term in the correct portion of the record where it should? Explain.
vi. Based upon what you learned, did this search run better or worse than before for each database? Why?

d. Let's modify this search again to: breast cancer[title] AND human[organism] and run the search at Entrez again and analyze the results as in exercise 4c above.
   i. Again, save your search results in MyNCBI.
   ii. As before, collect the data in your data table as to how the search ran in the Nucleotide, Protein, Gene and OMIM databases and analyze it.
   iii. Investigate where your "breast cancer" search term is located in the records at each database and add this information to your data table.
   iv. What is the best search to run at Entrez for the Nucleotide(core), Protein, Gene or OMIM databases? Why?
   v. What is the difference in running a [text] search, a [title] search limit or a search without limits at all, as you did for the first search in this module? Where is each searching in the database record and why does this make a difference in the search working better or not?
   vi. Is there such a thing as a perfect one-fits-all-search when using the Entrez search engine? Explain.

e. In running your searches now and looking over the differing results in each database, you should have realized that breast cancer in humans has a medical name (breast neoplasia). You will also have run across at least two different genes that cause breast cancer: BRCA1 and BRCA2. In bioinformatic databases, the names of genes involved in a particular disease can be substituted for the name of the disease in a search. There are other "bioinformatic synonyms" for breast cancer, but for this exercise, let's now make this search more specific to the terms typically used at Entrez by substituting the gene names for the concept of "breast cancer" in this manner: (brca1[title] OR brca2[title]) AND human[organism]. This search uses Boolean Logic. The parenthesis tells the search engine to run the search with the terms within any parentheses before it runs any other part of the search. The [title] tells the search engine to restrict it to the Title field of the records and the OR tells the search engine: the titles must have EITHER brca1 OR brca2 in them. That is, the search will return records with either brca1 or brca2 or both terms in the title. Once this search is done, the search engine will take the results and then AND them with the word human appearing only in the organismal field of the record. When you were using PubMed's Preview/Index Tab, this was created for you automatically. In Entrez you have to learn to do it manually yourself.
   i. Run the search and compare and analyze the results as you did for the searches in Module 4d. Would you say switching to using gene names instead of the disease name is an effective search strategy in Entrez? Why or why not?

ii. How would you modify the search above to include searching for these genes or their equivalent ones in the fruit fly *Drosophila melanogaster* model organism in addition to searching for human genes?

f. We finally have the search running the way we would like it to do so in Entrez for most of the databases. When we click through to the individual databases, it might be necessary to individually tailor the search for that particular database, but now we have an idea of which search string works best in what Entrez database and how to modify it.

g. Using the brca1/brca2 search results practice using the Preview/Index tab in the Entrez databases with the following exercises;
   i. Locate all the records at Nucleotides that contain only the complete coding sequence (CDS) for the brca1 gene.
   ii. Locate all the records at Protein that contain only a partial coding sequence for the brca2 gene.
   iii. Locate the records within Nucleotide that contain the exons for the brca1 and brca2 genes. What are "exons"? How many exons are there for brca1? For brca2?

h. Still using the results from the human brca1/2 search, perform the following exercises to learn what kinds of additional information you can learn by using the various hyperlinks available within Nucleotide or Protein records. Remember to use any of the NCBI Help documents you located in Module 1 as needed.
   i. Using the Preview/Index Tab, locate only those records for either brca1 or brca2 within Nucleotide that contain a hyperlink (a cross-reference) to the Human Genome Nomenclature Committee's (HGNC) resource for these genes. How many records exist for brca1? For brca2? Click through the link of any of the recovered records to HGNC. Are you still at NCBI? What is the information here and how would you make use of it?
   ii. Go to Entrez's Home Page and type in NM_000059.3 in the search box at the top of the page.
      1. How many results does Entrez return in what databases?
      2. Click on the Nucleotide result.
         a. What does NM_000059.3 mean?
         b. What is this record?
         c. What is GI:119395733 in the record and how does it differ from the NM record number above? Can it be used as a search term also?
         d. What kind of biological molecule is this?
         e. What does PRI mean at the top of the record?
         f. What are some synonyms for this gene that you could add as "OR" search terms if you wanted to recover more bioinformatic information about this molecule? Where in the record is this information found?
         g. What is the equivalent protein record at NCBI for this molecule? Where in the record is this information found?
         h. Click on the protein link—where does it take you?
         i. Print out this record and that of NM_000059.3. Compare and contrast what information is available in both. What is different and what is the same?
         j. Find the term "refseq" in the NM_000059.3 record. What is RefSeq and how does if differ from a GenBank record? (You will probably have to refer to the various help documents you have located earlier to answer this).

   k. Pull down the Display menu and select "Graph". What happens? How does this information relate to the text-version of the record?
   l. Find the check box at the top of the record that says "snp". Check it and then hit the "refresh" button. What information was added from what NCBI database to the record when you did this?
   m. Someplace in this record and in all Nucleotide records is a link that takes you to some information about the history of this record and all of the changes made to it since it was first added to NCBI. At the bottom of this page it tells you when the record was first "seen" at NCBI. What is that date?
   n. Finally, in the upper right hand corner of the record, there is a hyperlink called "Links". Click and hold down your mouse button on this. New options will appear. List the options available and then click through on PubMed. Is this all the literature that exists for this gene in PubMed? Develop a method to find out and test this. If it is not all the literature, then what does it represent?
   o. Click on the Gene link within the Links menu. What is Gene and what relationship does the Gene database have to the protein and nucleotide record you have been working with already?
   p. Where in the Gene record does "NM_000059.3" appear?
   q. From this exercise you should now know the NCBI accession number records for the gene, the mRNA and the protein for the breast cancer gene brca1. What are they?

5. Using the techniques above, develop similar "portraits of diseases and their genes" from the list below, using Entrez to search. The purpose of this writing and analysis exercise is to focus to learning what each separate Entrez database has in biological information that can be applied to a research project. DO NOT click through to other databases although you will be tempted to do so.
   a. Developing at least two additional search strategies than those used for the brca1 gene, making use of the various Help documents to guide you.
   b. Add new search terms to your search string as you discover them as Boolean "OR" statements.
   c. Read the titles and descriptions of the records you recover from at least the Nucleotide, Protein, Gene, OMIM, OMIA (if available) and PubChem (if available) databases. You may add other Entrez databases if you wish except for PubMed and PubMed Central.
   d. Develop "gene portraits" for humans and compare its function with at least one other organism by searching and modifying your searches as needed.
   e. What is the equivalent function of the gene/protein in another organism? Does it cause a disease?
   f. Save ALL your searches in a new MyNCBI with the password and username given to you by your instructor so that s/he can look over your searches for this project.
   g. Write your "portrait" as no more than a two-page, single-spaced paper, in New York Times Roman 10-12 point font. On the third page, copy/paste your searches and your successful search strategy from MyNCBI. On the same page, provide the Entrez databases and record accession numbers for the resources you used to create your portrait. Choose from one of the following diseases or conditions:
      i. Cystic fibrosis
      ii. Asthma

iii. Autosomal recessive/dominant Deafness
iv. Colon Cancer
v. Macular Dystrophy
vi. Marfan Syndrome
vii. Niemann-Pick Disease
viii. Neurofibromatosis
ix. Porphyria

Module 4: NCBI Entrez Gene—Discovering Gene Structure and Function Data

In the previous exercises you have learned to search both PubMed and NCBI Entrez. By now you should realize that there are many ways that NCBI presents information to researchers about genes and proteins. Many different records exist for the same gene or sequence. Each record contains some information, but not all the information available for a protein or gene. You have to learn to search well and then use the hyperlinks within records to travel around the NCBI site to obtain all the information you may need. In the last ten years or so, researchers have come to realize that that here is often no correct structure or sequence for a gene. One gene may encode more than one protein and/or RNA. There are "pseudogenes" sequences which are DNA sequences which share many of the features of true genes, but are not active and cannot produce a protein or RNA product. Even sequencing the same gene several times within the same laboratory results in different sequences due to sequencing errors or subtle distinctions in experimental techniques and/or software processing of sequence data. Many organisms have equivalent genes, but differing sequences. Regardless, NCBI collects ALL this information, creating a sequence record for each one. As a result, the same DNA, protein or RNA may have hundreds of records NCBI. Even the best search may recover 30-100 records about the same sequence. To help point researchers interested in gene structure and function, NCBI has created the Entrez Gene database where all "gene-centered" information is pared down to a single record for the gene, its sequence and structure. The Gene database links to all information about a gene's protein and RNA products, its function and similar genes and/or sequences in other organisms. Learning to run a search, locate the record that contains part of the information you need and then linking through to the Gene database immediately will get you to where you need to learn about a gene, its protein or its RNA products. Essentially, Entrez Gene is a "jump station" to locating more information very rapidly at NCBI.

There are two ways to link to Entrez Gene when looking for information at NCBI. One is to run searches as you already have in Entrez or in one of its databases specifically (e.g., Nucleotide or Protein). NCBI often calls this kind of search a "back door" search because you are linking into major jump station resources like Gene after searching another NCBI database. The other way is to search first in Entrez Gene if you already know what you are looking for. We will practice both here.

Back Door Searching for Gene Structure and Function Data. You want to study the major genes in humans responsible for colon cancer. Researchers, when they have this kind of research question, typically will choose to first search the Entrez Nucleotide database which is the major DNA sequence database at NCBI.

1. Go to Entrez (all databases) and without searching, click on the Nucleotide database. The reason for doing the search first in Nucleotide is you know you want to search just DNA sequences for these kinds of genes and because you will have use of the Details and Preview/Index Tabs within Nucleotide. This option does not exist if you were to globally search all the databases at Entrez simultaneously as we have done in the module before this one.

2. At Nucleotide, use the Preview/Index Tab to select for colon cancer in the title and humans in the organism fields. Run the search. How many total hits are there?
3. In the past year, NCBI has begun to add patent sequence data to Nucleotides. Using the Limits Tab, exclude any patent hits in your search. How many results do you have now?
4. Save this search in MyNCBI. Whenever you run the search again, it will automatically run the limits you set as well.
5. Go back to your search results in Step 3 and display the first 100 hits in a single web page.
6. Look over the titles of the records and make notes of the different kind of hits recovered. If necessary, open the links and look at the actual record. Record the different kinds of biological molecules this search recovered.
7. Now click on the "RefSeq" tab just above your first hit. How many hits do you have now? Record the different kinds of biological molecules here and compare them to those you found in step 5 above. What has happened?
8. With these data you have collected from this search and using any Help documents or searches in PubMed, write several paragraphs on the difference between RefSeq and the other Nucleotide records and why that difference helps to focus searches more effectively.
9. Looking over the hits under the RefSeq tab, list the record accession numbers only for those gene(s) involved in colon cancer and give a brief summary of each gene, including its gene symbol and what is known about it by clicking through to the actual Nucleotide record(s).
10. Choose one of the records from step 9 above and click into the actual record. Once at the record, you will find in the upper right hand corner a text link called "Links". Left click with your mouse on it and a pop-up menu will appear. Select Gene and click.
11. You are now at the NCBI Entrez Gene database for this gene, having migrated into it through a "back door" search first through the Entrez Nucleotide database. Compare and contrast the Gene and the Nucleotide database information for one of the human colon cancer genes.

Learning about Gene Structure and Function Data using Entrez Gene

Most researchers, after becoming familiar with a specific set of genes will search at Entrez Gene directly for the specific gene of interest.

1. Go to Entrez (all databases) and without searching, click on the Gene database.
2. Search for MLH1, one of the genes known to be involved in colon cancer in humans.
3. How many records are returned?
4. Display the first 100 in a single page. Look over the results, clicking through to the individual records, using your browser's "find in this page" command to locate where your search term, MLH1, is located and in what context. Write a short synopsis of how these records are related to each other in Gene and whether or not the MLH1 gene is found in other organisms. If so, which ones?.
5. Return to your full search for the MLH1 gene through the History Tab and restrict it to humans.
6. How many results are returned?
7. Again, return to your full search for MLH1 and now restrict it to a title search for MLH1 and the organism as humans.
8. Of the three searches, which works best and why?
9. Using Gene to learn about the structure and function of MLH1 requires you to learn to link to other sources continually. Look over this web page and note that there are links within the page, but also down the right hand side. All point to additional information about MLH1 and the same information may be accessible from more than one link.

10. Answer the following questions about MLH1. Some of the answers are directly on the MLH1 home page and others you will have to click through to new resources to answer. In some cases, you may have to link through to multiple resources to completely answer the question.
    a. What chromosome does MLH1 map to in humans?
    b. What other names (synonyms) is this gene known by?
    c. What is the equivalent gene in *E. coli*?
    d. What are the accession record numbers for its:
        i. mRNA?
            1. Click through and name the NCBI database that holds this record
        ii. Protein?
            1. Conserved domains for proteins are those sequences conserved through evolution and are very important in helping to determine the function of the gene and its protein. Clicking through if necessary, identify the conserved domains by both name and accession record number. What database at NCBI are they held in?
    e. What genes lie to the left (5') of MLH1 on the chromosome and what are their functions?
    f. What genes lie to the right (3') of MLH1 and what are their functions?
    g. What is a GeneRIF and what information does it provide about MLH1?
    h. What diseases are associated with MLH1 in humans?
    i. Where can you find ALL the literature published in PubMed about MLH1?
    j. What functions might the MLH1 gene/protein have in other organisms?
    k. Is the MLH1 gene expressed (makes protein) in other organisms besides humans?
    l. What potential RNAi's exist that could knock out the function of this gene to study it more? Where could you order them from?
    m. What cancer pathways is MLH1 involved in for humans?
    n. What mutations within the gene are known to cause a disease?
    o. What is the equivalent gene in mouse and on what chromosome in mouse is it on?
    p. The European Bioinformatics Institute (EBI) collects together all the information on the protein of a gene. It is an invaluable resource that complements what information NCBI collects on genes and their proteins. Find out how to migrate to EBI's Ensembl database from NCBI's Entrez Gene record for MLH1. Indicate how you do that, provide the accession number for the MLH1 Ensembl and list all the other organisms that also have a MLH1 protein similar to that found in humans.

Notes for the Instructor

This workshop was designed as a laboratory exercise for students, but it serves a dual role in acting as a tutorial for instructors integrating bioinformatic exercises and research into undergraduate biology laboratories. This section covers the necessary background information to plan and create bioinformatic instructional modules. The exercises are designed to be given in the order listed. Each module builds upon the latter. The modules contain search examples and build analytical skills needed to evaluate bioinformatic resources and "discover" what is available. Prior to actually searching the NCBI datasets, students first learn the interface and the various tools available within PubMed, upon which NCBI has built its sequence-related databases. Its interface, search process and results pages do not "violate expectations". Students are familiar with the process and purpose of searching and reading the published literature. The PubMed module is designed to visualize for students, those search strategies and process that are common in PubMed and the NCBI bioinformatic databases.

Because bioinformatics is still emerging as a method to perform research, the concept of searching to do research is foreign to students and probably to many instructors. Bioinformatics research is about data. It is presented to the researcher as search results with little context if any, and carries no intrinsic meaning or explanation as traditionally found in published research articles where the author provides the data in context within the additional information found in the Introduction, Results and Discussion sections of journal articles. Even the methodology in bioinformatic data sets is not typically presented to researchers. The entire concept of bioinformatics is to "discover" your own meaning by having the complete dataset available for your perusal and analysis. For this reason, instructing bioinformatics in laboratories requires a significant amount of background into both the science of the molecules at hand, the information science surrounding the bioinformatic resource and the development of analytical and critical thinking skills. A student studying genes needs to clearly understand both DNA and gene structure and function. What is not so obvious is that the same student needs to think informatically and understand the fundamental information architecture and function of any bioinformatic database being used. A lecture component to bioinformatic laboratories is almost a mandate, to insure that students have a context from which to work. For lower level undergraduates developing modules focused to a gene, disease, or condition, that has been showcased in the news headlines can provide the context needed. For most students, permitting them to choose a disease/protein/gene that carries a particular significance to that student works well, but instructors may find themselves scrambling to identify (and learn) the appropriate NCBI bioinformatic information needed for the student-initiated project for themselves. A more formal approach, and one that would fit more effectively into an existing curriculum, is to provide a context to any companion lecture course to the laboratory sections. Thus when introducing the concept of a gene in freshman courses, or studying the regulation of a gene in higher level courses, a bioinformatic laboratory module focused to working with NCBI's Gene resource in the former case or to identifying sequences that are genes and then downloading the promotor sequence in the latter case, would be appropriate. This approach would provide the students not only with a context to work within, but would extend the lecture with a practical hands-on research experience.

Suggestions on How to Prevent Students from Becoming Lost in "Biospace"

Because the intent of bioinformatics research is to "discover" what one needs to know, bioinformatic records have become highly interlinked within a given website and across the breath of the Internet to permit an almost infinite free association of ideas and data. Deliberately, no defined

"investigative" path or "protocol" exists for researchers or students. There is no information on what components are available to examine or how they are connected to each other. Even within the NCBI resources, students can become easily "lost" in that bioinformatic resource of over 70 different databases and tools. One of the ways the authors have found in teaching information resources in general, and almost a requirement for teaching bioinformatics, is to not only create the actual research project for the course, but to create a "toolbox" of resources from which the students are to work and to inform them that these resources and only these resources contain the information needed for the exercise or module. The toolbox should also contain resources and tools that would NOT necessarily contain the information for the exercise set. To help students learn to restrain themselves from hyperlinking too far into "biospace", it is also effective to have them keep a "bioinformatic research notebook" of where they found what they needed and the process by which they located the resource. The process of taking notes slows down hyperlinking and creates introspection.

A related tip is for students to use a web browser that supports tabbed browsing. Students should be instructed on how to use the tabs and to learn to "right click" with the mouse to open a link in a new tab. Bioinformatic data is often created dynamically by the server delivering the information. It is literally being projected to the computer screen. It is not being linked through a specific URL that is available in the browser's cache or history file. When the student clicks off that resource to investigate a possible lead to more information, the first link is lost completely in the browser's history cache. The student will not be able to back up and retrace her steps.

Finally, there is no uniform or regular standard that permits bioinformatic data or any particular resource to be saved. NCBI now does offer a Save Search feature for some, but not all of their resources, which is covered in this workshop. Most researchers and students learn to save web pages as an .html file or some will convert it to a PDF file if they have the proper software. One of the most common problems encountered for researchers and students, is that they simply forget where on their hard drive they saved their search results, web pages and other material for a given day or project. Developing computer file management strategies prior to teaching bioinformatics is a must. There should be a standard procedure given to students on how and where to save their various files: Where to place the files on the hard drive, in what hierarchical order, and what to name them. Instructions should be given to the class on the exact steps to back that material up on both the hard drive and external to their computer in at least two different format types. A typical back-up would involve saving the files to another folder (or better, a server) to the laboratory computer and then to download it to CD-ROMs or thumb drives at least. Because most bioinformatic resources are in the public domain, students can (and will) access these resources and work on their projects outside of the formal laboratory sessions. Some thought has to be given to synchronizing their files and results together when they access bioinformatic resources from different geographically-located computers.

General Overview of the Major International Bioinformatic Sequences Spaces

Understanding how bioinformatic information is organized worldwide is an important component of both teaching and acquiring bioinformatic research skills. Although this module deals exclusively with NCBI as a starting point for learning bioinformatics, the resources at NCBI are not all inclusive. There are two other major sequence spaces internationally and thousands of smaller, more highly purposed, "secondary" and "tertiary" databases, throughout the Internet. The major sequence spaces all start with the same core DNA or protein sequence, but organize and manage the information differently. This is a graphic representation of "informatics", which is defined as the organization and management of information. In turn, the smaller sequence spaces that are highly purposed, such as

FlyBase or the Protein Data Bank (PDB), draw their original sequence and related information from one of the major international bioinformatic sources. Conceivably, once a student has become comfortable working within one of these sequence spaces (such as NCBI), an advanced set of exercises would be to seek the same information elsewhere and compare and contrast each major space and/or secondary database with the other by working the same exercise in each resource.

The three major international biosequence spaces currently are NCBI (http://www.ncbi.nlm.nih.gov/, see below), the European Bioinformatic Institute (EBI, http://www.ebi.ac.uk/) and the National Institute of Genetics (NIG, http://www.nig.ac.jp/index-e.html). Researchers in the Americas upload their sequence data and related information to NCBI; those in the Pacific Rim upload their data to NIG, and European researchers upload to EBI. Once a day, NCBI, EBI and NIG upload new data to each other, update their files and organize and integrate the new sequence data into their respective sites and make it available to the public (Figure 1). Although the foundation of NCBI, EBI and NIG is uniformly the same sequence data, and the files sequence records they create have to share the same core set of features to make them interoperable to all researchers, each entity re-organizes the data and adds information to fit their particular needs, missions and cultural perspectives. EBI has always organized bioinformatic information around the concept of a protein and its function. Their emphasis is on human curation more than computerized curation. As a result, information for a given record at EBI is much more stable and less likely to change, as the information displayed is that for which the scientific community has developed a consensus opinion. They typically mount a single resource for each protein. NIG has always been more technologically oriented with an emphasis placed on data analyses and applications. A wide spectrum of computationally intense resources, including microarray data manipulation, or experimental sequence similarity analysis tools, are more likely to be found at NIG than elsewhere. NIG files also make use of text information much more than graphical displays as is found for EBI and NCBI. This can work well for some students, but not for others who need to process visual information to learn. This entire process of sharing bioinformation internationally is under the oversight of the International Nucleotide Sequence Database Collaboration (http://www.insdc.org/), including the creation of a common set of core sequence record features (http://www.insdc.org/files/documents/feature_table.html).

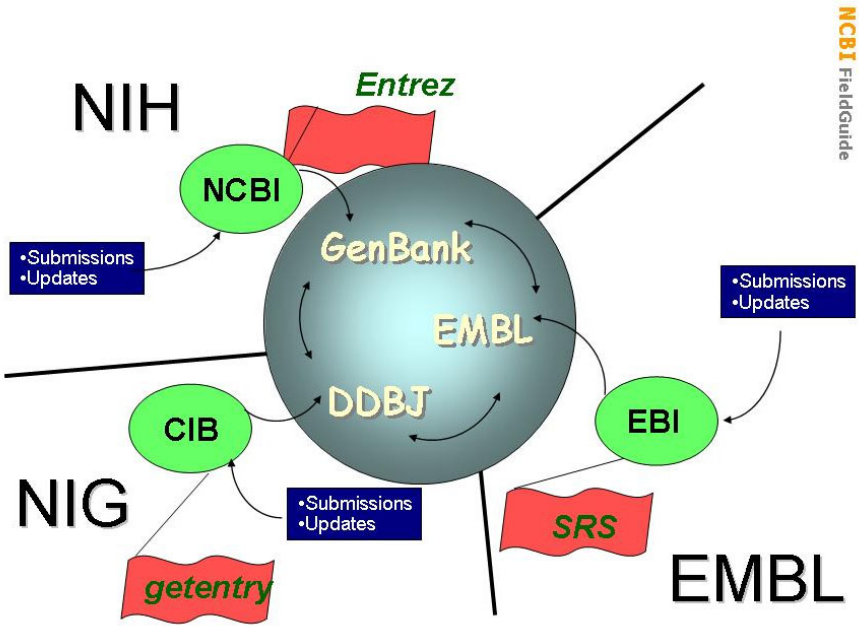

Figure 1. Data uploaded every day to one of the three world-wide "biospaces" (NCBI, EBI or NIG is shared among the three once a day. Each reorganizes the information to its own purpose, making the data available in considerably different formats and interfaces. Each biospace has its own equivalent search engine as shown here: Entrez for NCBI, getentry for NIG and the Sequence Retrieval System (SRS) for EBI.

Module 1: Learning Basic NCBI Web Site Navigational Skills

NCBI began as the GenBank database repository for the DNA sequence data arising from the Human Genome Project in the 1990's. The data from the Human Genome Project began as a trickle of information that rapidly exploded into an avalanche in the late 1990's. As a result, NCBI was building databases and linking them together faster than the staff could create the necessary help documents and organize them into one collated space. As a result, NCBI grew into a collection of databases extensively hyperlinked to each other in a vast three-dimensional "biosequence space" that contains over a billion separate records of information. NCBI has various site maps and one web site search engine to aid researchers in locating databases and tools from their home web page. In addition, each database or tool often has its own help documents and research aids located internal to NCBI at the home web page for each of its databases. Since each database at NCBI has a different set of data within it, its own interface and its own purpose, each is unique. To effectively use NCBI's resources, students must be able to locate the various help documents and tutorials related to each resource. This first module is to provide that experience. The ten questions in this module are designed to force the students to develop "survival" skills at NCBI, including learning how to migrate through NCBI. Additional, more specific navigational modules should be added by laboratory instructors depending upon the content and purpose of each individual bioinformatic laboratory "experiment" created. A bioinformatic laboratory dealing with human disease, for example, should have a navigational learning module constructed for the disease resources at NCBI: OMIM, OMIA, Malaria, MapViewer, dbSNP, etc., in addition to the exercise given here.

The NCBI Data Model and Its Significance to Teaching

As a unit of the National Library of Medicine at NIH, the role of NCBI is to develop public bioinformatic (molecular biology) resources that increase our understanding of diseases and health overall. It began in 1988 as the GenBank repository for raw DNA sequencing data arising from the Human Genome Project. Today it holds information for over 160,000 organisms, along with 3D protein structures, genomic mapping information, chromosomal information, diseases, PubMed literature and more. In contrast to EBI which organizes information around a protein sequence, NCBI is informatically DNA-centric, organizing information around the original GenBank sequence data uploaded by researchers for any given gene. It is also simultaneously disease oriented to fit its mission as defined by federal mandate.

As new protein and gene information began to pore out of the early DNA sequencing efforts, NCBI had to find a way to integrate these new data functionally into its sequence space. Essentially, NCBI had to re-define itself structurally. No longer could it be a single GenBank database holding sequence information. Much like designing a city, NCBI had to create an infrastructure and housing for sequence information while simultaneously building into it access to all of its components and defined "traffic patterns" of information flow that were consistent and immutable. To do this, NCBI developed the NCBI Data Model (Ostell, *et al*., 2001; Figure 2).

Much like the animal model systems as represented by *Drosophila* and *Caenorhabditis*, the NCBI Data Model provides a model database schema or framework in which to link and interlink new information to an existing sequence data record without altering the latter. The Data Model follows the Central Dogma, with information existing in separate, but hyperlinked, file records as related to DNA, RNA and protein *sequence* data. Thus, DNA, RNA and protein raw sequence data are uploaded into separate files, even if they came from the same research group. In turn, DNA, protein or RNA files are then collated into the appropriate database resources: Nucleotide, Protein or RefSeq, respectively. When,

for example, a protein is identified to a particular mRNA and/or DNA sequence, the appropriate records are linked together. Most of the NCBI sequence records are mounted automatically with computers, due to the large amount of sequence records continually uploaded to NCBI. As a result, compared to other bioinformatic resources, NCBI records are sparsely annotated. Each individual database resource provides partial information on a gene or protein, typically in a disease or pathological perspective to match the mission of NIH. Regardless of what information a particular flat file record contains and for what type of macromolecule, the NCBI Data Model dictates that each will point back eventually to one piece of information: the GenBank sequence. In addition, if a disease or pathology has been associated with a sequence, ultimately the flat file records linking the information for that sequence together, will also point back to that disease or pathology.

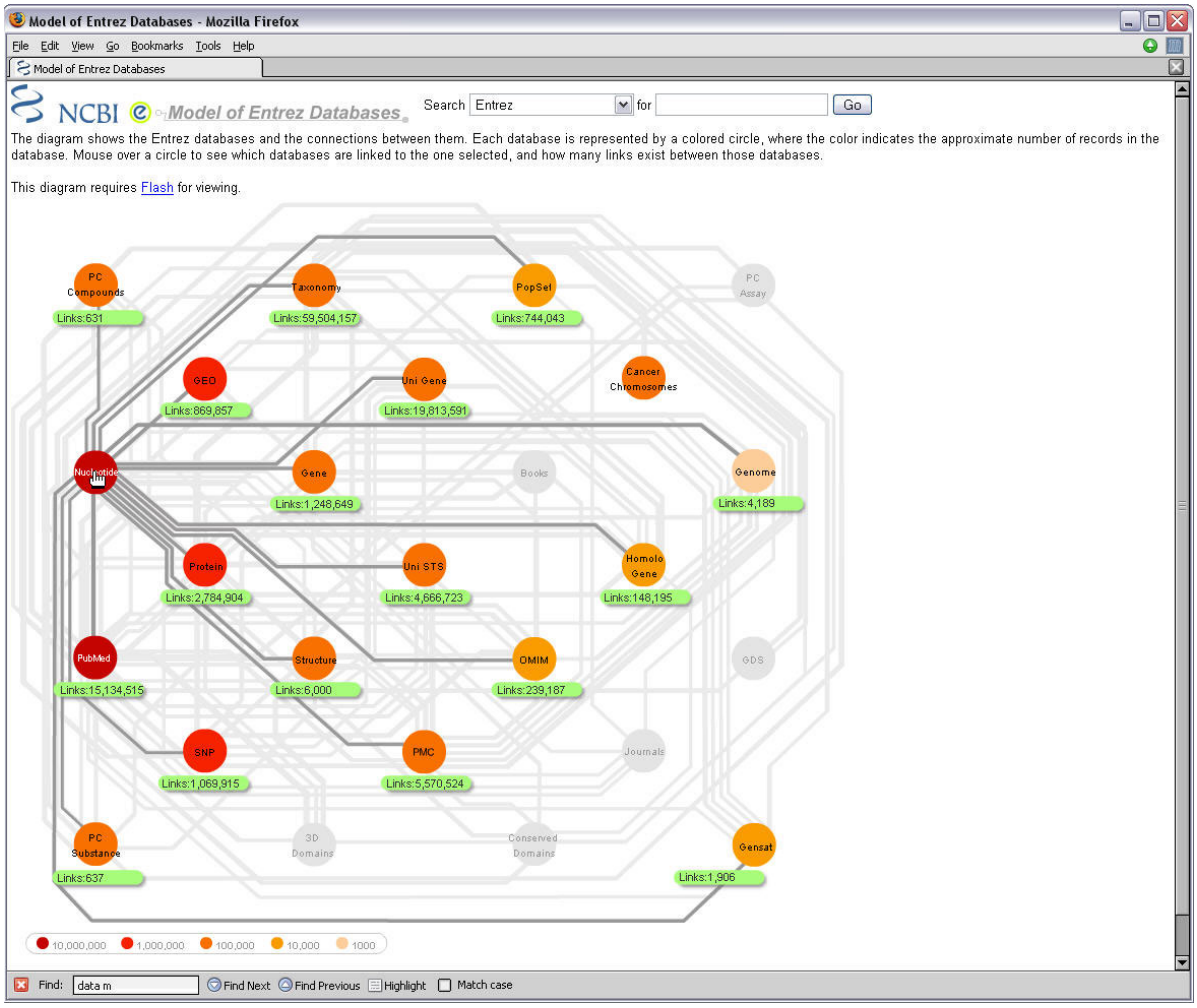

Figure 2. An interactive graphical view of the NCBI Data Model (http://www.ncbi.nlm.nih.gov/Database/datamodel/index.html). which sets the specifications for what databases will link to each other at NCBI and where within each record linking should occur.

As the emphasis in research shifted from DNA to protein and now to RNA and genomes, NCBI has been able to handle the new information and reorganize existing data by created newly linked datasets that contained entirely different forms of data that normally would be incompatible with each other in a single resource. The system is highly flexible and enables NCBI to rapidly mount new data to the public. It is the NCBI Data Model that permits users to link through and gather information concerning a particular DNA sequence from the PubMed literature, RNA resources and disease and protein information, as shown in Figure 3.

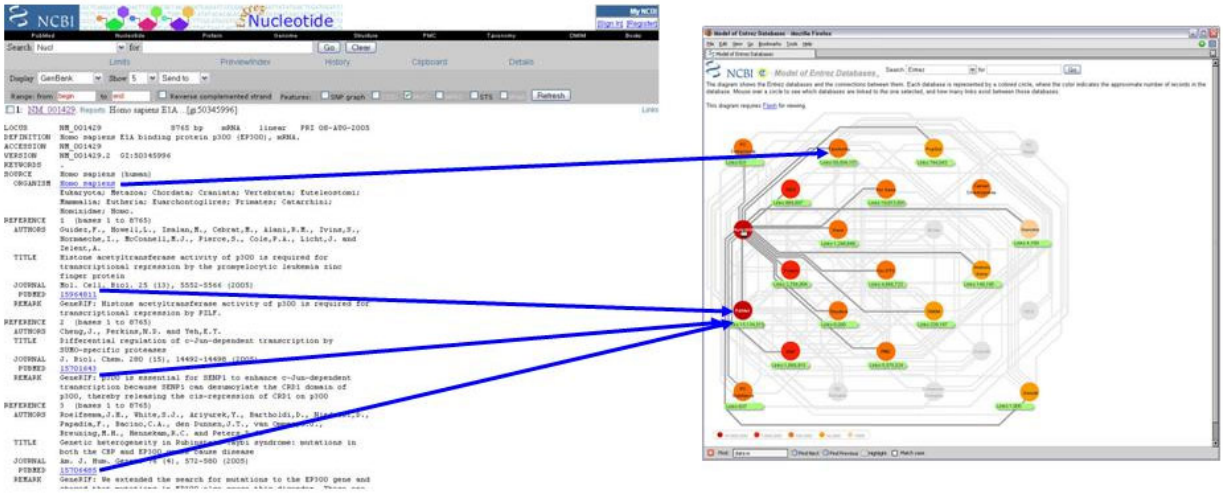

Figure 3. The NCBI Data Model specifies where in the flat file record hard links should occur between the NCBI databases. Each link on the Model represents a specific kind of link within the flat file.

Module 2: PubMed--Learning the NCBI Information Architecture the Easy Way

Before there was bioinformatics at NCBI, there was the PubMed database. When it came time to build GenBank to hold the human genome data, the scientific community began searching for the appropriate place to house the information. The National Library of Medicine at NIH was chose due to its PubMed expertise. The database was the most advanced and robust online database at the time. It was publicly available and the budding NCBI division that had built it had by far the most experience and expertise in building online resources. GenBank was created by NCBI with the PubMed model in mind. To this day, many of the NCBI databases still retain the original information architecture of PubMed: The Details, Limit, Preview/Index, History and Clipboard tabs, along with the concept of varying the display, "sending" results to be saved, emailed or to the clipboard, etc. (Figure 4). Obviously the information underneath these tabs will vary dependent upon the database. Even those databases at NCBI that do not model the PubMed interface exactly still build upon such concepts as "find related" and "links". The former are searches based upon a computational algorithm that find similar concepts in the MeSH headings, title and abstract text fields for each search run, while the latter links out to related resources both internal and external to NCBI. Learning to use PubMed is essential to being able to expertly search NCBI resources. This is particularly important when creating instructional modules for laboratory exercises by instructors.

Figure 4. The common functional interface shared by many of NCBI's databases. Learning the functions of the various Tabs in one database enables searching skills in another. The best place to begin instruction into these functionalities is with PubMed since most students will already understand the scope, purpose and results of searching literature databases.

Module 4: NCBI Entrez Overview—Searching for Bioinformatic Data

Currently, the NCBI web site contains two different kinds of research resources:

1. Biological (bioinformatic) databases that hold data, curated and annotated to varying degrees dependent upon the resource, some examples of which are:
2. Bioinformatic tools.
   - BLAST (http://www.ncbi.nlm.nih.gov/BLAST/), NCBI's sequence similarity searching set of tools
   - MapViewer (http://www.ncbi.nlm.nih.gov/mapview/, a chromosome browser for any of the 800 completed genomes at NCBI)
   - Cn3D (http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml, a web-based 3D protein sequence viewer),
   - Electronic PCR (http://www.ncbi.nlm.nih.gov/sutils/e-pcr/)
   - RefSeq (http://www.ncbi.nlm.nih.gov/RefSeq/, a collection of reference sequences for genomic DNA, transcript RNAs and proteins
   - A mixed bag of other informational tools such as online books (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Books).

Due to the many types of NCBI data and resources, NCBI has created a series of different search portals to locate these different kinds of databases and tools at NCBI. These portals are not actually databases although their interfaces often look as if they are. Each search portal is designed to return a list of links to a specific subset of NCBI databases that are functionally interrelated. Each of NCBI's search

portals is based upon an entirely different kind of search based upon the type of information need and currently include.

- Entrez: Text-based searching of the flat file records across 33 separate databases highly hyperlinked to each other. It does no search for data, such as sequence data.
- BLAST: Searches specifically for sequence data within a NCBI record. The search string is a sequence that is copy/pasted into the search box. BLAST then searches for similar sequences by mathematically calculating the probability that the search string sequence is similar to any of the sequences available in NCBI records (http://www.ncbi.nlm.nih.gov/BLAST/). There are five general forms of BLAST, a BLAST program for each organism for which NCBI holds its complete genome sequence, and highly specialized BLAST programs such as comparing two sequences directly. NOT an Entrez resource, BLAST results will hyperlink back to Entrez databases.
- Gene: Gene is both a component of Entrez and a text-based search engine highly specific for gene-related information. Not all genes are indexed in the Gene resource (discussed in the Gene module below). However, because Gene is composed of RefSeq sequence records specifically (also discussed later), Gene is one of the few ways to retrieve ONLY RefSeq records.
- PubMed: Also a member of the Entrez suite, PubMed is the familiar text-based search engine configured specifically to search the research literature. It is extensively hyperlinked to the bioinformatic records at NCBI and to all of the 33 databases within NCBI Entrez.
- MapViewer. Essentially a chromosomal browser, MapViewer searches and displays genomic, gene, transcript and disease information by chromosomal position (http://www.ncbi.nlm.nih.gov/mapview/). Researchers can choose from many different types of sequence, cytogenetic and radiation hybrid maps. NOT an Entrez database, it will link back to Entrez resources.
- Taxonomy Browser: Also an Entrez resource, Taxonomy is simultaneously a browser/search tool for taxonomic information by organismal name, either common or scientific (http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/). It also contains Entrez summary record data for each organism including number of gene, protein, transcript sequences at NCBI and links to external resources for each organism.
- VAST: A 2D/3D protein search engine (http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml). The input search is a set of 3D protein spatial coordinates from a VAST record. Essentially this search engine finds protein sequences that would take the same shape and/or hold the same dimension in space, regardless of the specific protein sequence.

PubMed has already been covered in Module 2. The rest of this article deals with the two other major search portals, Entrez and Gene, both of which are easily mastered by undergraduates with limited biology knowledge.

If one could view the biological databases and their organization at NCBI, one would see a nucleus of 33 bioinformatic databases that are highly interlinked to each and cross-searchable in a resource called "Entrez" (Figure 5). The resources within Entrez range from proteins and nucleotide sequence information (Protein and Nucleotide databases respectively) to literature (PubMed), 3D structures (Structure), single nucleotide polymorphisms (dbSNP) and much more. Researchers enter search terms and Entrez globally searches for the term across its 33 resources. It then depicts the results in a graphical table (Figure 6). Clicking on the database of interest within the results table takes the researcher to the actual resource itself, much like when searching PubMed takes you to the full text of the article or abstract directly.

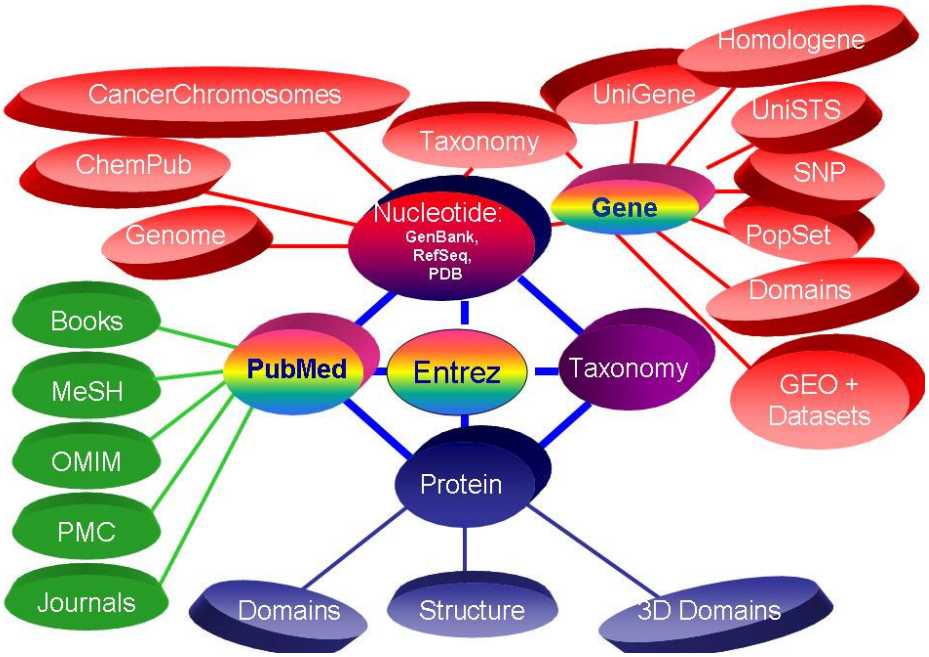

Figure 5. A graphical representation of what a portion of the NCBI Entrez suite of databases could look like if it could be seen. Searchers finding information in one database may have to migrate through several resources to locate information. Although it might not seem so when searching, the hyperlinks are positioned in a highly structured architecture.

Figure 6. The home page of NCBI's Entrez suite of 33 databases in 2007. Clicking on each database will take you to that database's specific home page to search. Alternatively, one can globally search all 33 databases at once as shown here, although the search may not necessarily run in each database as intended. Entrez cues searchers to this through the use of grey boxes. In those databases, the search ran but was permuted.

Hyperlinking within Entrez is highly structured. The search process across the suite of Entrez databases occurs in a specific pathway. Entrez first searches a core of four resources: PubMed, Nucleotide, Protein and Taxonomy. If it finds a hit in any of these resources, it retains the results and then moves on to search those sub-databases associated with that original core database searched. Therefore, Entrez is a text-based search engine at NCBI that retrieves links to individual data records from within the Entrez nucleus of resources and displays them as an "index" of links. The most important factor in searching Entrez is to understand that it only searches 33 of the greater than 70 resources at NCBI. However, Entrez has its own idiosyncrasies, which are quite distinct from what most researchers expect from searching PubMed. Learning to use Entrez effectively developing appropriate search strategies that Entrez can interpret and knowing which database resource within the results table contains the data one is seeking. Module 3 takes students through this process in a parallel manner in which they learned to restrict searches in PubMed to yield more relevant and focused results. Unlike PubMed, however, Entrez neither automatically maps the search to a standard vocabulary nor does it automatically build a Boolean search string. Researchers must switch to these skills to make use of the NCBI biological databases. The module here provides a basic approach to this. Providing lectures on Boolean logic as part of the laboratory experience would enable more sophisticated searches than the example given here.

Outside of PubMed, THE most important component of NCBI's bioinformatic databases may well be the "Properties" menu under the Preview/Index Tab of the various Entrez databases. Each database will have different options under these tabs. The student module 3 introduces this by having students search the Preview/Index tabs at Entrez for specific features. The module deliberately reduces the instructions to the students, who now have to "discover" just where in the Preview/Index Tab's various options the information needed can be found (primarily the Properties and Text options). This structured set of exercises mimics the typical discovery process by researchers at bioinformatic databases who have to develop strategies to seek what may be at a bioinformatic site in the way of the data and information then need with little or no framework or help documents to provide the way.

However, having a thorough knowledge of the Properties option of the Preview/Index Tab is also critical for instructors building laboratory modules. For example, you want to create a bioinformatic module on *Drosophila*. NCBI holds many records on *Drosophila*, including from the third-party resource, FlyBase. To locate which DNA sequence records of *Drosophila* exist at NCBI, you must use the Preview/Index Tab, choose "properties". Unfortunately, the choices you most often will see will be in a software programming set of terms with which you will be largely unfamiliar. For example, there are various "dbxref" options. This represents a hyperlink somewhere in a NCBI record; a cross-reference link to another database. By choosing "dbxref flybase", you can locate all the records at NCBI in any given database that are from FlyBase and upon which you can build instructional modules. The way you locate this information is to perform a search at the NCBI Web Site pull-down menu for "dbxref". Alternatively, if one wanted to create an instructional module concerning *Drosophila* to complement a laboratory exercise on the Mendelian genetic segregation pattern of *Drosophila* eye color, through the Preview/Index Tab, you could build the following search at the Nucleotide database: **"**drosophila melanogaster"[Organism] AND eye[Title], recovering the 10 DNA sequence records available within the Nucleotide (DNA) sequence database within minutes. Thus the same skills that are taught for students in this article can be used by instructors in a twist that meets their professional pedagogical needs.

NCBI's Flat File Record Format and Its Significance to Teaching

The NCBI Data Model described above specifies how each record must be created at NCBI and in what format. For example, each must have an accession number, be linked back to any earlier records, have a "features" section, a "reference" section, list the sequence at the end of the record, and depending upon the particular kind of record and molecule, structure or organism, what additional fields should be placed into the record and to what other records it should be linked. Sequence flat file records at NCBI are now too extensive to show in print here. However, NCBI has created a teaching version of a general GenBank record which can be accessed at this URL directly http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html (Figure 7) or through the "Alphabetical List" link available at the top left-hand navigational panel on its web pages.


NCBI Sample GenBank Record

PubMed Entrez BLAST OMIM Taxonomy

GenBank Flat File Format

Click on any link in this sample record to see a detailed description of that data element or field. All of the descriptions are included on this page, so it can be printed as a single document. You can also return to the Alphabetical Quicklinks Table or Resource Guide

```
LOCUS       SCU49845     5028 bp    DNA             PLN       21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
  TITLE     Cloning and sequence of REV7, a gene whose function is required for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
  PUBMED    7871890
REFERENCE   2  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast requires Axl2p, a novel
            plasma membrane glycoprotein
  JOURNAL   Genes Dev. 10 (7), 777-793 (1996)
  PUBMED    8846915
REFERENCE   3  (bases 1 to 5028)
  AUTHORS   Roemer,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-FEB-1996) Terry Roemer, Biology, Yale University, New
            Haven, CT, USA
FEATURES             Location/Qualifiers
     source          1..5028
                     /organism="Saccharomyces cerevisiae"
                     /db_xref="taxon:4932"
                     /chromosome="IX"
                     /map="9"
     CDS             <1..206
                     /codon_start=3
                     /product="TCP1-beta"
```

Figure 7. The GenBank Sample record is a true "live" record at NCBI for which NCBI persistently retains a teaching version. Clicking on any of the blue text in the Sample Record takes the searcher to an explanation of that particular component of the flat file.

The NCBI GenBank Sample Record and its related resources are an invaluable resource for creating teaching modules in the basic modes of migration through bioinformatic resources. Learning to read a bioinformatic database record, being able to ascertain what information exists in it and how to get to more information through hyperlinking is very different than reading a journal article record, yet it is a necessary form of scholarly communication in the biosciences today. Without being able to understand

the structure and format of NCBI's database "space" and what file formats it holds where, students will not have the basic fundamental skills to discover the biological information that resides in bioinformatic resources and put it to use, whether at NCBI or elsewhere.

NCBI maintains the sample GenBank Flat File record permanently with a persistent URL. It is not the true "live" record of this sequence. Instead, down the left-hand side of the record are blue text hyperlinks will take the reader to an explanation of that particular component of the sample record and/or additional help files. In the "live record" any text links on the left hand usually take the reader to another bioinformatic database or resource. There are many different kinds of flat file records in GenBank; the most simplistic being those that contain DNA or protein information. However, what is known about a particular molecule dictates the structure of the record and exactly what fields of information exist. For example, some records will have promoter, TATA, 3- or 5'-untranslated fields, or alternatively spliced fields in the record. Others will not. Thus the actual fields and information available for a particular molecule in its specific flat file record at NCBI is highly variable and unique to each molecule. The GenBank Sample Record is generically representative of a fairly simple NCBI gene record. As a result, NCBI has provided links at the bottom of the GenBank Sample Record to other "live" records which contain additional fields. As for the GenBank Sample Record, NCBI maintains these additional records permanently at the same URL so that they are always available to both educators and students as teaching/learning tools. Finally, NCBI carries a link within the Sample Record to its own "live" record. To access these additional records, you must scroll down the Sample Record until you locate this information.

Module 4: NCBI Entrez Gene—Discovering Gene Structure and Function Data

In the past, NCBI linked internally to its own databases. But as more information is collected about genes, proteins and diseases, NCBI is now linking out of its flat files to centralized "link index" resources like Gene at NCBI, and from their on to additional sources and databases both internal and external to itself. Students have to learn to migrate from linked resource to resource, analyze the information and continuing linking through to additional resources to build a complete "portrait" of knowledge about a gene or a protein. Unfortunately, learning to migrate between the various NCBI resources is becoming more and more complex and confusing, as not all of NCBI's resources are necessarily linked to each other directly. To get to a needed resource from one database, users must often travel through other resources. Module 4 instructs this through the use of NCBI's Entrez Gene resource. In information science, Gene is considered an "index" resource. By itself, it contains little if any original information. Instead, in a cohesive manner, Gene organizes gene-centric information within NCBI in a consistent manner with just enough textual information to make sense before providing a series of hyperlinks to the original source of the information. In order to use Gene effectively, students (and instructors) must be familiar with the scope and intent of the resources to which Gene will ultimately link.

Not all genes are represented in Entrez Gene, which collects only RefSeq records for display. Instead of lecturing on the distinction between a RefSeq record and a typical Protein or Nucleotide database record, a component of the Gene Module has students investigate this directly. Explaining the concept of RefSeq, a database of "non-redundant" reference sequences for which NCBI's staff, through a set protocol, will choose one representative record/gene/organism, is often difficult to visualize until students actually can see the difference in search results. Thus if one is searching for the human MLH1 in RefSeq, one will retrieve only one record. If searching in Protein or Nucleotide, searchers will

retrieve every single bit of sequence information loaded up to NCBI. Hundreds of overlapping sequence records with information from all over the world for each gene exists within NCBI's GenBank and Protein databases. From this milieu, NCBI will curate one single representative record for a gene/protein instead. However, this process of determining the representative record/gene/organism is time consuming. Thus RefSeq does not yet represent every single gene known for each organism. One of the values of learning to use Entrez Gene is to learn effective and efficient discovery of gene-related information by learning to properly seek and hyperlink to new information. However, one of the caveats is that Gene only collects RefSeq records to make that possible and, thus, does not represent all genes sequenced at this time. The answers to Question 10 of the Gene Module 4 are provided below.

Answers to Question 10 of the Gene module on MLH1

Question: Answer the following questions about MLH1. Some of the answers are directly on the MLH1 home page and others you will have to click through to new resources to answer. In some cases, you may have to link through to multiple resources to completely answer the question.

a. What chromosome does MLH1 map to in humans?
   ANSWER: Chromosome 3
b. What other names (synonyms) is this gene known by?
   ANSWER: COCA2, FCC2, HNPCC, HNPCC2, MGC5172, hMLH1
c. What is the equivalent gene in *E. coli*?
   ANSWER: MutL
d. What is the accession record number for its mRNA and the NCBI database that holds this record?
   ANSWERS: NM_000249 and Nucleotide
   What is the accession record number for its protein product?
   ANSWER: NP_000240
e. Conserved domains for proteins are those sequences conserved through evolution and are very important in helping to determine the function of the gene and its protein. Click through if necessary and identify the conserved domains by both name and accession record number. What database at NCBI are they held in?
   ANSWERS: HTPase_C, MutL_Trans_MLH1, MutL; accession numbers cd00075.3, cd03483.1, and COG0323.2, respectively from the Conserved Domains Database (CDD). The information can be found from the pop-up menu that appears if one clicks on the NP record accession number in the Genomic regions area of the flat file or uses the CDD link from the right-hand list of links.
f. What genes lie to the left (5') and right (3') of MLH1 on the chromosome and what are their functions?
   ANSWER: This answer is found in the Genomics Context area of the Gene flat file record.
g. What is a GeneRIF and what information does it provide about MLH1?
   ANSWER: This answer is found in the Bibliography area of the Gene flat file record.
h. What diseases are associated with MLH1 in humans?
   ANSWER: This answer requires one to link through from multiple places within the MLH1 Gene record, including the Books, OMIM and GeneTests links in the right-hand link list and the "additional links" at the bottom of the record. For more advanced students, essentially migrating to PharmGKB, MGC and through HGNC and LinksOut to additional tertiary records, will build an advanced knowledge of diseases and phenotypes associated with this gene.
i. Where can you find ALL the literature published in PubMed about MLH1?

ANSWER: PubMed link in the right-hand links list. PubMed (GeneRIF) only finds those articles in PubMed used to build the Gene, Nucleotide and Protein records at NCBI.

j. What functions might the MLH1 gene/protein have in other organisms?
   ANSWER: Use the HomoloGene link in the right-hand Links list.

k. Is the MLH1 gene expressed (makes protein) in other organisms besides humans?
   ANSWER: Use the UniGene link in the right-hand Links List.

l. What potential RNAi's exist that could knock out the function of this gene to study it more? Where could I order them from?
   ANSWER: Use the Probe link in the right-hand Links list.

m. What cancer pathways is MLH1 involved in for humans?
   ANSWER: Use the KEGG (and could use PharmGKB for tertiary database information) from the right-hand Links list.

n. What mutations within the gene are known to cause a disease?
   ANSWER: There are a variety of ways to answer this question. The full answer will come from using the OMIM and SNP links from the right-hand Links list, as well as clicking through PharmGKB and HGNC to other resources.

o. What is the equivalent gene in mouse and on what chromosome in mouse is it on?
   ANSWER: You must click through the HGNC list and then secondarily to the MGI (Mouse Genome Informatics) to get this answer easily. This will take you directly to the equivalent mouse Gene record at MGI.

p. The European Bioinformatics Institute (EBI) collects together all the information on the protein of a gene. It is an invaluable resource that complements the information NCBI collects on genes and their proteins. Find out how to migrate to EBI's Ensembl database from NCBI's Entrez Gene record for MLH1. Indicate how you do that, provide the accession number for the MLH1 Ensembl and list all the other organisms that also have a MLH1 protein similar to that found in humans.
   ANSWER: Again, the most direct way is to click through to HGNC from the right-hand Links list and then on from there to Ensembl. HGNC is essentially a bioinformatic record link resolver for several major bioinformatic sources. When in one of those resources, if you click through HGNC, it will take you directly to the equivalent record in another resource elsewhere.

Materials

Either a Windows XP or Mac OS 10.x computer connected to the Internet with a broad-band LAN connection (not wireless) is required. Ideally, each student should have his/her own computer, but working in pairs can be effective if students share time working at the computer. Either the most recent version of Internet Explorer or Firefox browsers is required for Windows XP while Firefox is strongly suggested for Mac computers. Regardless of browser, it should be Java-enabled which typically is automatically loaded when the above browsers are downloaded, installed and launched for the first time.

Literature Cited

Almeida, C.A., D.F. Tardiff, J.P. DeLuca, and M.F. Hall. 2003. Using bioinformatic software to understand the Central Dogma of biology. Journal of College Biology Teaching, 29:15-23.

Bednarski, A.E., S.C.R. 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, 4:207-220.

Honts, J.E. 2003. Evolving strategies for the incorporation of bioinformatics within the undergraduate cell biology curriculum. Cell Biology Education, 2:233-247.

Krawetz, S.A. 2000. Design and implementation of an introductory course for computer applications in molecular biology and genetics. Methods in Enzymology 125:449-460.

Kumar, A. 2005. Teaching systems biology: An active-learning approach. An active-learning approach. Cell Biology Education, 4:323-329.

Mulnix, A.B. 2003. Investigations of protein structure and function using the scientific literature: An assignment for an undergraduate cell physiology course. Cell Biology Education 2:248-255.

Ostell, J.M., S.J. Wheelan, and J.A. Kans. 2001. The NCBI Data Model (Chapter 2). Pages 19-43, *in* Bioinformatics: a practical guide to the analysis of genes and proteins. John Wiley & Sons, 470 pages.

Ranganathan, S. 2005. Bioinformatics education—Perspectives and challenges. PLoS Computational Biology, 1:447-448.

Rice, M., W. Gladstone and M.Weir. 2004. Relational databases: A transparent framework for encouraging biology students to think informatically. Cell Biology Education, 2:231-252.

Weaver, T. and S. Cooper. 2005. Exploring protein function and evolution using free online bioinformatic tools. Biochemistry and Molecular Biology Education, 33:319-322.

Wefer, S.H. 2003. Name that gene. The American Biology Teacher, 65:610-613.

About the Authors

Diane Rein received her Ph.D. degree in Developmental Biology from the Institute for Developmental Research at the Cincinnati Children's Hospital Medical Center in 1977. As a postdoctoral fellow at the Stanford Medical Center and Baylor College of Medicine, she extended her early interest in neuroblastoma non-muscle contractile proteins to the cellular, molecular and genetic study of human muscle diseases, primarily muscular dystrophy. Subsequently, she held an active faculty position at the University of Cincinnati where she taught undergraduate and graduate courses in the biological sciences and directed a federally funded graduate research program in mammalian DNA polymerase-beta excision repair processes that involved characterizing both the proteins and genes involved. Beginning in 1994, Diane became active as an analyst/consultant for various biotechnology companies. It was during this time that she became fascinated in scientific information transfer that ultimately led to her obtaining a Masters degree in Library and Information Science from the University of Illinois, Urbana-Champaign in 2001. Diane currently is an EduCollab Project member of the National Center for Biotechnology Information (NCBI) which develops and provides advanced bioinformatic instruction for information specialists and is one of the instructors for the Introduction to Molecular Biology Resources continuing-education 3-day course sponsored jointly through the National Library of Medicine, the Medical Library Association and NCBI.

Jennifer Sharkey is Assistant Professor of Library Science and Information Integration Librarian at Purdue University Libraries. Sharkey's research areas include 21st Century teaching methodologies, technology integration, and utilizing design principles (graphic, instructional, web) for curricula development. Recent publications include the article "Towards Information Fluency: Applying a Different Model to an Information Literacy Credit Course" and the book chapter "Beyond the Keyboard: Optimizing Technology Spaces for Collaborative Learning, Instruction, and Service" in Teaching with Technology: An Academic Librarian's Guide. She is a member of the American Library Association and Association of College and Research Libraries. For more information, visit: http://www.lib.purdue.edu/ugrl/staff/sharkey/eportfolio/.

Jane Kinkus is the Mathematical Sciences/General Sciences Librarian and Assistant Professor at Purdue University. She holds a Bachelor of Science joint degree in Mathematics and Philosophy and a Master of Library Science degree from the University of Pittsburgh. In addition to her general interest in helping college students learn critical thinking and information literacy skills, Jane's research interests include project management in libraries, and the use of digital technologies in the library/academic environment.