Course materials relating to this workshop are posted on the internet at the following URL: http://www.bioc.rice.edu/~olson/courses.html

II. Rules of thumb about bioinformatics exercises for beginners:

Advantages:
a. Most schools have computer labs or library computer access.
b. Exercises only require a browser.
c. Students interested in all areas of bioscience must learn to mine computational databases - unavoidable for future scientists.

Pitfalls to avoid:
d. Databases are HUGE, must give specific, step-by-step instructions or the student will get hopelessly lost.
e. Some sites go down temporarily, design a 2- or 3-part exercise that uses more than one site, or make the due date flexible.
f. Practice the exercise each year, databases and interfaces change rapidly - what worked last semester may not work now.

III. URLs or google searches for beginning bioinformatics students:
A. The Human Genome Project:
http://www.ornl.gov/sci/techresources/Human_Genome/project/about.shtml

Chosen to emphasize the importance of computers in biology, questions are designed to illustrate the overwhelming amount of data generated by genome sequencing centers. Storing the data is problematic in and of itself, but analyzing and comparing that data is completely impossible without the help of the computers.

The student can answer basic questions about genome sequencing, such as:
  i. How many genes are found in the human genome?
  ii. How many DNA base pairs make up the human genome?
  iii. Why are computers essential to genome sequencing centers? Name 2 project goals of the Human Genome Project that require the help of computers.
  iv. What kind of instructions do genes encode?
  v. What functions do proteins perform?
B. NCBI Databases and related journal articles

The National Center for Biotechnology Information (NCBI) is a web site belonging to the National Institute of Health and the National Library of Medicine. The NCBI website and selected article from the journal Nucleic Acids Research stress the size and growth rate of biological databases. By reading the article about GenBank, students can answer questions about growth and typical users of this database, such as:

i. How many new sequences were added in 2007?
ii. What are the top 5 species with sequences in GenBank?
iii. What does a GenBank entry include?
iv. Who typically submits data to GenBank?
v. What website would you use to retrieve GenBank information?

C. Search Engines and Search Design

PubMed provides access to citations from a broad range of scientific literature. When searching this or any database, the terms used are important. General terms can return thousands of hits. For example, try searching for articles about hemoglobin. How can you identify which of those articles relate to your topic? Adding additional terms to narrow your search can help. However, search terms that are too specific may mean that relevant articles are missed.

D. The Human Genome Sequencing Center at Baylor College of Medicine
http://www.hgsc.bcm.tmc.edu/

The Houston Medical Center is the largest in the world, and students are generally interested to know that important medical research is performed locally. Questions can be designed that guide the student through an exploration of the web site:

i. Is the Houston Medical Center involved in genome sequencing research?
ii. How many genome projects are being sequenced for different organisms at the Human Genome Sequencing Center, Baylor College of Medicine?
iii. Choose one of the links to the organism project home page that interests you most and view the picture of the organism. Select one of the tabs at the top of the page to view another organism. All of these genomes have either been sequenced, or are currently in the process of being sequenced.
iv. How many primate genome projects are listed?
v. Why do you think so many primate genomes are being sequenced?
vi. Why is it important to humans to learn about bovine genomes?
vii. Why is it important to humans to learn about microbial genomes?
E. For more advanced exercises, the following websites have bioinformatics tools where students can perform sequence comparisons, BLAST searches for similar sequences in a database, multiple sequence alignments and many different types of computational sequence analyses.

The Biology Workbench
http://workbench.sdsc.edu/

National Center for Biotechnology Information:

Expasy (Swiss Institute of Bioinformatics)
http://us.expasy.org/tools/

European Bioinformatics Institute
http://www.ebi.ac.uk/Tools/

Rice Connexions – search for Bios 533 Bioinformatics Course
http://cnx.org/

IV. Sequence Alignment.
Simple BLAST search - healthy vs. diseased sequence comparison: our example is beta Hb S (sickle cell) vs. beta Hb A (normal)
Type the following web site for the NCBI into the browser:

Next to the “Search” box, select Protein, to search the NCBI database containing protein sequences.

In the query box next to “Search Protein for”, type 1HBSB, then click “Go”.

The record for hemoglobin S should be returned. Hemoglobin is the protein in our blood cells that carries oxygen. Click on the link entitled “1HBSB”.

Now you are looking at the record for chain B of the human hemoglobin S protein.
Next to the word Display in the grey region at the top of the file, change “GenPept” to “FASTA”. This will display the amino acid sequence for hemoglobin S in FASTA format.

Hold down the left mouse button while you move the mouse over the sequence. This should highlight the amino acid sequence in blue. Now choose “Edit:Copy” from the browser window, or hit the buttons “Ctrl” and “C” to copy.
Now, click on the NCBI logo in the upper left corner of the web page to return to the main page.

In the dark blue menu bar at the top of the page, click on the word “BLAST”.

In the list of options, click on the link entitled “protein blast”.

Click in the Search box and choose “Edit: paste” from the browser menu or hit the “Ctrl” and “P” keys to paste the sequence into the search box.

Change the “nr” database to “swissprot”, then click the BLAST! button.

Wait for the search to complete. The browser will display the longest sequence alignments represented by colored bars.

Under the graph indicating the length of the top alignments, there will be a list of aligning sequences in order of decreasing alignment scores. Click on the score of the first item in the list, which is the highest scoring alignment. This will take you to the section of the file where you can view the alignment.

Identify any differences in the sequence of Query 1 and Subject 2

It turns out, the difference in the two sequences is the sickle cell mutation in Hemoglobin. Sickle cell anemia is a blood condition seen most commonly in people of African ancestry and in the tribal peoples of India. The individual must have two copies of the mutant hemoglobin gene to exhibit the sickle-shaped cells indicative of the condition. The sickle cell mutation reflects a single change in the amino acid sequence of the beta subunit of hemoglobin, the protein that is the main oxygen-carrier in the blood. The beta subunit has the amino acid valine at position 6 instead of the glutamic acid that is normally present. This alteration is the basis of sickle cell disease.

V. Multiple Sequence Alignment and Tree Generation.
Exercise in Multiple Sequence Alignment: - molecular relatedness in the same protein family within one organism: our example is representative sequences from the globin-like superfamily

The Biology Workbench
http://workbench.sdsc.edu/

The Biology Workbench is one of my favorite teaching tools, because the student can do a complete Bioinformatics project with the Workbench, from retrieving the sequences to performing multiple alignments and creating phylogenetic tree diagrams.
Step 1 is register; the email authorizing the account is usually sent out virtually immediately, there is no significant time involved in requesting a new account.

Step 2 is to retrieve sequences under the Protein Tools Menu. They can be uploaded, but the simplest way to retrieve them is to use the Multiple Database Search Tool (Ndjinn) within the Workbench. Be careful, there are many databases - too much information - too many results from a query confuses the student. It’s easiest to recommend one database for the student to search (the instructor should definitely work through any exercises in advance to determine the best set of written directions for the student).

Enter the workbench, and scroll down the page until the five menu buttons are visible. The "Session Tools" button allows the naming of a session, so that different jobs in progress can be saved under distinct sessions. Select "Session Tools", then select "Start New Session" and click on "Run" to change the name of "Default Session" to a new name. Once the workbench has been exited, the session will remain. Subsequently, clicking on the dot to the left of the session name under the "Session Tools" menu, and then selecting "Resume Session", will recall the session. The Workbench policy at the time of this writing is that old jobs are deleted only when an account has not been accessed for 6 months.

Next, select "Protein Tools" from the menu buttons, highlight "Ndjinn Multiple Database Search", and click "Run". In the query box to the right of the term "Contains", enter the search string indicated in the table at the end of the syllabus. Scroll down the database list and check the box to the left of the database entitled "SWISSPROT" before hitting the "Search" button. Direct the student to import the records by checking in the box to the left and hitting “import sequences”. Continue searching and importing sequences as above. Some sequences listed in the table will need to be entered using the "Add New Protein Sequence" tool. Copy and paste sequences into the provided space and click “Save” to import. Once all sequences are imported, go to step 3.

Step 3 is to perform multiple alignments with the ClustalW tool under the Protein Tools Menu. This will generate multiple alignments (the alignment must be exported to be saved, which will allow access with the alignment tools later), rooted and/or unrooted trees and pairwise alignment statistics. Check the boxes to the left of each of the desired sequences. Scroll down the protein tools menu and highlight "CLUSTALW - Multiple Sequence Alignment", then click "Run". The default parameters will be sufficient for our purposes, so on the next page just select "Submit". When the sequence alignment is returned, scroll down the page and view the multiple alignment. The alignment can be copied and pasted into a text document for reports. The Workbench automatically returns an unrooted tree with the alignment. Look at the unrooted tree. Ask the student questions that require them to analyze the alignments and the tree:
Problem 1
Which two sequences appear to be most closely related by viewing the alignment and the unrooted tree?

Problem 2
Which are most distantly related?

Problem 3
Look at the multiple sequence alignment scoring section. Notice the sequence list that assigns numbers to each sequence. The alignment scores are labeled by the assigned sequence numbers, so this list is necessary to interpret the scores. According to the pairwise scores, which two sequences are most similar?

Problem 4
What is the score of the best pairwise alignment?

Return to the top of the page and select "Import Alignment(s)". This will save this sequence alignment under the "Alignment Tools" menu of this session.

VI. Shading Alignments and Rooted Phylogenetic Trees
This optional step involves viewing the multiple alignment, editing it or generating figures for reports and publications, all of which can be performed using selections found under the Alignment Tools Menu. Select an alignment from part V by clicking in the small box to the left of the listing.

1. Under the Alignment Tools menu, check the box next to the desired alignment. Select “DRAWGRAM – Draw Rooted Phylogenetic Tree from Alignment” and click “Run”. The default settings are appropriate, so click “Submit” on the next page. This returns a rooted phylogenetic tree image. Click the link “Download a Postscript version of the output” to obtain a copy with white background suitable for printed reports. Ask the students: Do the same two sequences appear to be the most closely related by viewing the rooted tree in comparison to the unrooted tree that was produced with the alignment?

2. Open the alignment in one of the text editors in Alignment Tools, such as TEXSHADE or BOXSHADE (note the warning under BOXSHADE that when the ruler is chosen for alignment numbering, boxshade can get stuck and never finish- the ruler is a nice addition to the figure, so try changing the font size or page orientation when this happens, as the Workbench recommends). The student can edit the alignment manually or change the coloring scheme of the alignment through these editors. They can use them to generate a nice multiple alignment figure for a paper.
<table>
<thead>
<tr>
<th>Protein</th>
<th>Search string to locate protein sequence using Ndjinn search of the swissprot database in Biology Workbench</th>
<th>Search string to locate protein sequence using NCBI's Protein search of the swissprot database and enter sequence manually into Biology Workbench</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Human Hemoglobins</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HBB</td>
<td>hbb human hemoglobin; import the protein sequence titled &quot;HBB_HUMAN&quot;</td>
<td></td>
</tr>
<tr>
<td>HBS</td>
<td></td>
<td>1HBSB</td>
</tr>
<tr>
<td>HBG1</td>
<td>fetal hemoglobin; import the protein sequence &quot;HBG1_HUMAN&quot;</td>
<td></td>
</tr>
<tr>
<td>HBG2</td>
<td>fetal hemoglobin; import the protein sequence &quot;HBG2_HUMAN&quot;</td>
<td></td>
</tr>
<tr>
<td>HBE</td>
<td>human hemoglobin epsilon; import the protein sequence &quot;HBE_HUMAN&quot;</td>
<td></td>
</tr>
<tr>
<td><strong>Leghemoglobins</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Soybean</td>
<td>lgba soybean</td>
<td></td>
</tr>
<tr>
<td>Pea</td>
<td>lgb4 pea</td>
<td></td>
</tr>
<tr>
<td>Corn</td>
<td>hbl maize hbm</td>
<td></td>
</tr>
<tr>
<td>Rice</td>
<td>hbl1 orysj</td>
<td></td>
</tr>
<tr>
<td><strong>Truncated Globins</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2-on-2 from Arabidopsis</td>
<td>* Perform Ndjinn search of GENPEPT database using 110738756</td>
<td>AAK55409</td>
</tr>
<tr>
<td>S. aureus protein</td>
<td></td>
<td>NP_371527</td>
</tr>
<tr>
<td>truncated globin from <em>M. tuberculosis</em></td>
<td>glbn myctu</td>
<td></td>
</tr>
<tr>
<td><strong>Sensor Globins</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HemAT from <em>B. subtilis</em></td>
<td>hemat bacsu</td>
<td></td>
</tr>
<tr>
<td>HemAT from <em>H. salinarum</em></td>
<td>hmat halsa</td>
<td></td>
</tr>
<tr>
<td>methyl-accepting chemotaxis protein from <em>A. tumefaciens</em></td>
<td></td>
<td>NP_354049</td>
</tr>
</tbody>
</table>