GenBank and Protein Data

4/9/2008

Opening Discussion

- Do you have any questions about the assignment?
- Do you have any questions about the reading?

GenBank

- We have already looked at the site for GenBank a bit.
- GenBank also defines flat text file formats that specify the information for a sequence.
- Being able to parse these files is a valuable skill.
- Chapter 10 goes into detail on different approaches for pulling information out of GenBank files.

Multiline RegEx

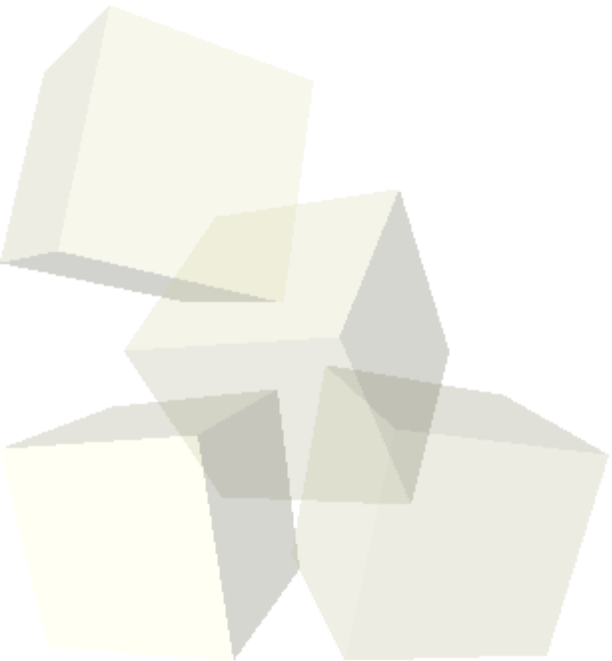
- Many of the files you will deal with have multiple lines in them. There are two options that you can use to help deal with these types of strings.
- The //s option will make it so that . can match newline.
- The //m option makes it so that ^ and $ work not just for the beginning and end of the string as a whole, but also for beginning and end of each line.

Flags and Loops

- One common approach for dealing with GenBank or other large files is to read in lines as part of a loop and keep a flag (or flags) to give you information about whether you have reached critical parts of the file.
- For example, a GenBank file has the word “ORIGIN” right before the sequence data.
- Let's see who we can use this technique.

DBM Files

- DBM files are like simple databases. They basically tie a hash to a data file.
- You use the command dbmopen to connect a has with a data file. It can take three arguments: the hash, a file name, and the permissions in octal.
- When you are done using the DBM use dbmclose to sever the tie and close off the file.

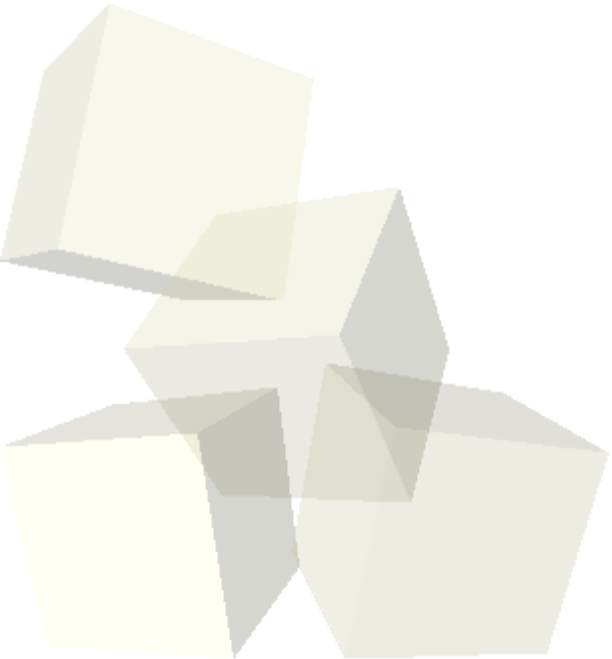


PDB

- Another significant source of bioinformatic data is the Protein Data Bank. (http://www.rcsb.org/pdb/)
- This site has directories of information that you can download to get protein structures and other information.

Dealing with Directories

- PDB data doesn't come in a single file. It comes in a directory structure.
- For this reason, we need to know how to deal with directories.
- The opendir subroutine will open a directory and readdir will read in the names of the files and directories. When you are done, use closedir.



Closing Remarks

- Remember that there is no class on Friday.
- Try to get me the other submission of project #1 soon.
- Assignment #7 is due today. Assignment #8 is due on Monday.