Lab #2
BioPython

BioPython

- Sequence format conversion
- Sequence maniplulation
- Interface to common programs/databases
  - BLAST, Clustalw, EMBOSS, SCOP, SwissProt, ...
- PubMed & Medline access
- Simple GUI programs
- BioSQL integration

- http://biopython.org/DIST/docs/tutorial/Tutorial.html

Simple SwissProt Example

```
from Bio import ExPASy
from Bio import SeqIO
handle = ExPASy.get_sprot_raw("A0LR17")
seq_record = SeqIO.read(handle, "swiss")
handle.close()
```

Note: help() comes in handy here…

BLAST

```
from Bio.Blast import NCBIWWW
from Bio.Blast import NCBIXML

result=NCBIWWW.qblast("blastp","swissprot",
"MAKMIAMADEAARRALERGMNQLADAVKVTLGPKGRNVVLEK
KWGAPTITNDGVSIAKEIELEDPYEKIGAELVKEVAKK")
blast_record = NCBIXML.read(result)
result.close()
blast_record.alignments
```

Pubmed

```
from Bio import Entrez
from Bio import Medline

# Always tell NCBI who you are
Entrez.email = "sludtke@bcm.edu"
handle = Entrez.esearch(db="pubmed", term="Ludtke SJ[Author]",
retmax=500)
record = Entrez.read(handle)
print record
ids=record["IdList"]

handle=Entrez.efetch(db="pubmed",id="22696402",rettype="medline")
records = list(Medline.parse(handle))
```

Medline Terms

Affiliation [AD]	Investigator [IR]	Pharmacological Action [PA]
Article Identifier [AID]	ISBN [ISBN]	Place of Publication [PL]
All Fields [ALL]	Issue [IP]	PMID [PMID]
Author [AU]	Journal [TA]	Publisher [PUBN]
Author Identifier [AUID]	Language [LA]	Publication Date [DP]
Book [book]	Last Author [LASTAU]	Publication Type [PT]
Comment Corrections	Location ID [LID]	Secondary Source ID [SI]
Corporate Author [CN]	MeSH Date [MHDA]	Subset [SB]
Create Date [CRDT]	MeSH Major Topic [MAJR]	Supplementary Concept[NM]
Completion Date [DCOM]	MeSH Subheadings [SH]	Text Words [TW]
EC/RN Number [RN]	MeSH Terms [MH]	Title [TI]
Editor [ED]	Modification Date [LR]	Title/Abstract [TIAB]
Entrez Date [EDAT]	NLM Unique ID [JID]	Transliterated Title [TT]
Filter [FILTER]	Other Term [OT]	UID [PMID]
First Author Name [1AU]	Owner	Version
Full Author Name [FAU]	Pagination [PG]	Volume [VI]
Full Investigator Name [FIR]	Personal Name as Subject [PS]	
Grant Number [GR]		

Lab #2

- Download the starter program from the class website. The starter program queries PubMed for articles written by a specific author, and prints the number of publications and number of coauthors. Try the program and understand what it's doing Use this program as a starting point for the following exercise. Each person in the group should select and complete one task (any of the 4), then one person should combine all of the work into a single program, which should be emailed to the TA before the end of class. You should coordinate your data representations so your code will all work together. Make sure each person in the group understands how all of the components work before leaving. As with lab 1, the tasks are ordered from most difficult to easiest:

1. Rather than simply counting the number authors in the retrieved records, tabulate the number of times each author name appears, then print a) the number of authors and b) the top 10 authors with their publication count.

2. Similar to 1), tabulate all of the words present in the title of all of the retrieved records. Eliminate common words like "the" and "and", and print an ordered list of the top 20 occurring words.

3. Very often it is impossible to identify authors uniquely by their PubMed names. "Ludtke SJ" could be "Ludtke S". "Chen F" could represent dozens of different authors. While there is no perfect solution to this problem, to help assess the impact of the first issue, count the number of unique AU values, and also count the number of unique family names. Print the difference between these counts as the number of possible overlaps.

4. Modify the program to query based on a provided keyword rather than an author's name. Pubmed does not like overly large queries, so be sure to restrict the number of retrieved records to a reasonable number (no more than 1000). Be sure to print a warning if this limit was reached.