Welcome to BST 281 Lab 7

22 Mar, 2018

Mike MacArthur

macarthur@g.harvard.edu

Office Hours: Fridays 2-3p

Kresge Student Lounge

Assignment 4 now posted

Due March 30th by 11:59pm

If your terminal is no longer working in Atom, go to Settings>Packages and update the platformio package

---

Parsing GFF files

What are GFF files?

General feature format files

Used for annotating features in biological sequences

There are multiple formats (GFF2, GFF3, GTF) that each have defined structures

Get the lab9_2.csv file

Before running, `pip install bcbio-gff`

```
import pprint
from BCBio.GFF import GFFExaminer
from BCBio import GFF
import csv

in_file = "SRS013876.with_fasta.gff3"
examiner = GFFExaminer()

in_handle = open(in_file)
pprint.pprint(examiner.parent_child_map(in_handle))
in_handle.close()

in_handle = open(in_file)
recList = []
for rec in GFF.parse(in_handle):
   recList.append(rec)
in_handle.close()

print(recList[0])
```

GFF Files can also be handled as delimited file use the csv package

Practice problem:

Write a script that opens the GFF file and writes each row to a new csv file

Hint: you'll probably want to use csv.reader() and csv.writer()

The answer is in the Lab9_Answer.py script

---

grep

grep is a command line tool that allows you to search for specific text, or patterns of text in a designated file

Usage: `grep [OPTIONS] PATTERN [FILE]`

To better understand how grep works, lets look at a simple home-made Python version

Download the grep.py script and she.txt file

```
def grep(word, filename):
    """Implement unix command grep.
    The grep command takes a string
    and a file as arguments and prints
    all lines in the file which contain
    the specified string.

    $ she.txt
    I'm sure that the shells are seashore shells.
    So if she sells seashells on the seashore,
    The shells that she sells are seashells I'm sure.
    She sells seashells on the seashore;

    $ python grep.py sure she.txt
    The shells that she sells are seashells I'm sure.
    I'm sure that the shells are seashore shells.

    """
    lines = open(filename).readlines()
    return [line for line in lines if word in line]

if __name__=="__main__":
    import sys
    word, filename = sys.argv[1:3]
    print("".join(grep(word, filename)))
```

What exactly is the script doing?

How do you pass the pattern and file parameters to the function?

Let's run the script. On the command line run

`python grep.py sure she.txt`

What is returned?

You can alter this function, or use a function like this one to search for rows in your GFF file that contain certain strings...

Performing the same grep operation using the command line utility

`grep sure she.txt`

Questions or optional continue working on the Jupyter Notebook from lecture