Algorithmic Approaches for Biological Data, Lecture #14

Katherine St. John

City University of New York
American Museum of Natural History

9 March 2016

Outline

- Project

Outline

- Project
- Scraping Data from the Web: `urllib` and friends

Outline

- Project
- Scraping Data from the Web: `urllib` and friends
- Parsing Structured Data: genbank & SQL

Outline

- Project
- Scraping Data from the Web: `urllib` and friends
- Parsing Structured Data: genbank & SQL
- Counting Revisited: Using Dictionaries to accumulate pattern counts; Scraping data from the web

Project

- Idea: apply what you learned to your thesis or research.

Project

- Idea: apply what you learned to your thesis or research.
- Include at least three techniques from the class.

Project

- Idea: apply what you learned to your thesis or research.
- Include at least three techniques from the class.
- Send an abstract of what you would like to do and what data sets and methods it will entail.

Project

- Idea: apply what you learned to your thesis or research.
- Include at least three techniques from the class.
- Send an abstract of what you would like to do and what data sets and methods it will entail.
- Next: timeline, two milestones, final report & presentations on the last day of class.

Project

- Much data is stored in a very structured way (e.g. CSV, FASTA, PNG)

Project

- Much data is stored in a very structured way (e.g. CSV, FASTA, PNG)
- Already seen: FASTA files:

Project

- Much data is stored in a very structured way (e.g. CSV, FASTA, PNG)
- Already seen: FASTA files:
  - Names (identifiers) occur on lines that begin with ">".

Project

- Much data is stored in a very structured way (e.g. CSV, FASTA, PNG)
- Already seen: FASTA files:
  - ► Names (identifiers) occur on lines that begin with ">".
  - ► Sequences follow the name, often on multiple lines, if long.

Project

- Much data is stored in a very structured way (e.g. CSV, FASTA, PNG)
- Already seen: FASTA files:
    - Names (identifiers) occur on lines that begin with ">".
    - Sequences follow the name, often on multiple lines, if long.
- HyperText Markup Language (HTML): formatting commands for webpages.

Project

- Much data is stored in a very structured way (e.g. CSV, FASTA, PNG)

- Already seen: FASTA files:

  - ▶ Names (identifiers) occur on lines that begin with ">".
  - ▶ Sequences follow the name, often on multiple lines, if long.

- HyperText Markup Language (HTML): formatting commands for webpages.
  Demo: show the raw file of rggs seminar page

Project

- Much data is stored in a very structured way (e.g. CSV, FASTA, PNG)
- Already seen: FASTA files:
    - Names (identifiers) occur on lines that begin with ">".
    - Sequences follow the name, often on multiple lines, if long.
- HyperText Markup Language (HTML): formatting commands for webpages.
  Demo: show the raw file of rggs seminar page
- National Center for Biotechnology Information (NCBI) databases:

Project

- Much data is stored in a very structured way (e.g. CSV, FASTA, PNG)

- Already seen: FASTA files:
    - ▶ Names (identifiers) occur on lines that begin with ">".
    - ▶ Sequences follow the name, often on multiple lines, if long.

- HyperText Markup Language (HTML): formatting commands for webpages.
  Demo: show the raw file of rggs seminar page

- National Center for Biotechnology Information (NCBI) databases:
  diverse collection ranging from raw data to publication indexing service (PubMed).

urllib2 Commands

```
</tr><tr class="bgdark"><td valign="top" style="vertical-align: top">  <a name="eztoc810013_0_1_0_1_2" id="eztoc810013_0_1_0_1_2"></a><h6>Spring Term 2016</h
  </td>
</tr><tr class="bglight"><td valign="top" style="vertical-align: top">   
  </td>
</tr><tr class="bgdark"><td valign="top" style="vertical-align: top">  <a name="eztoc810013_0_1_0_1_3" id="eztoc810013_0_1_0_1_3"></a><h6>25-Jan-16</h6>
  </td>
</tr><tr class="bglight"><td valign="top" style="vertical-align: top">  Matthew Aardema, American Museum of Natural History
  </td>
</tr><tr class="bgdark"><td valign="top" style="vertical-align: top">  "Natural selection in Lepidoptera: biological studies across scales&quot;
  </td>
</tr><tr class="bglight"><td valign="top" style="vertical-align: top">   
  </td>
</tr><tr class="bgdark"><td valign="top" style="vertical-align: top">  <a name="eztoc810013_0_1_0_1_4" id="eztoc810013_0_1_0_1_4"></a><h6>01-Feb-16</h6>
  </td>
</tr><tr class="bglight"><td valign="top" style="vertical-align: top">  Siobain Duffy, Rutgers University
  </td>
</tr><tr class="bgdark"><td valign="top" style="vertical-align: top">  "Begomovirus evolution: fast, biased and taxonomically out of control&quot;
  </td>
</tr><tr class="bglight"><td valign="top" style="vertical-align: top">   
```

urllib2 Commands

- Including the module:

urllib2 Commands

- Including the module: `import urllib2`

urllib2 Commands

- Including the module: `import urllib2`
- Opening a file:

urllib2 Commands

- Including the module: `import urllib2`
- Opening a file:

```
webpage =

urllib2.urlopen("http://www.amnh.org/our-research/richard-gilder-graduate-school/academics-and-research/
```

urllib2 Commands

- Including the module: `import urllib2`
- Opening a file:

  ```
  webpage =
  urllib2.urlopen("http://www.amnh.org/our-research/richard-gilder-graduate-school/academics-and-research/
  ```

- Reading from a file:

urllib2 Commands

- Including the module: `import urllib2`

- Opening a file:

  ```
  webpage =

  urllib2.urlopen("http://www.amnh.org/our-research/richard-gilder-graduate-school/academics-and-research/
  ```

- Reading from a file:
  - `webpage.read()`: reads the entire file into a single string.

urllib2 Commands

- Including the module: `import urllib2`
- Opening a file:

  ```
  webpage =
  urllib2.urlopen("http://www.amnh.org/our-research/richard-gilder-graduate-school/academics-and-research/
  ```

- Reading from a file:
  - `webpage.read()`: reads the entire file into a single string.
  - `webpage.readline()`: read the next line of the file.

urllib2 Commands

- Including the module: `import urllib2`
- Opening a file:

  ```
  webpage =

  urllib2.urlopen("http://www.amnh.org/our-research/richard-gilder-graduate-school/academics-and-research/
  ```

- Reading from a file:
  - `webpage.read()`: reads the entire file into a single string.
  - `webpage.readline()`: read the next line of the file.
  - `webpage.readlines()`: read the file into a list of strings.

urllib2 Commands

- Including the module: `import urllib2`
- Opening a file:

  ```
  webpage =

  urllib2.urlopen("http://www.amnh.org/our-research/richard-gilder-graduate-school/academics-and-research/
  ```

- Reading from a file:
  - `webpage.read()`: reads the entire file into a single string.
  - `webpage.readline()`: read the next line of the file.
  - `webpage.readlines()`: read the file into a list of strings.

- Closing a file:

urllib2 Commands

- Including the module: `import urllib2`
- Opening a file:

  ```
  webpage =
  ```

  ```
  urllib2.urlopen("http://www.amnh.org/our-research/richard-gilder-graduate-school/academics-and-research/
  ```

- Reading from a file:
  - `webpage.read()`: reads the entire file into a single string.
  - `webpage.readline()`: read the next line of the file.
  - `webpage.readlines()`: read the file into a list of strings.

- Closing a file: `webpage.close()`

bioPython

- A module for biologists (based on bioPerl).

bioPython

- A module for biologists (based on bioPerl).
  bioPython page

bioPython

- A module for biologists (based on bioPerl).
  bioPython page
- Need to install:
  `conda install biopython`

bioPython

- A module for biologists (based on bioPerl).
  bioPython page

- Need to install:
  `conda install biopython`

- Many useful functions but will focus on sequence I/O and interactions with NCBI (Entrez).

bioPython

- A module for biologists (based on bioPerl).
  bioPython page

- Need to install:
  `conda install biopython`

- Many useful functions but will focus on sequence I/O and interactions with NCBI (Entrez).
  ```
  from Bio import Entrez
  #NCBI requires an email from each user:
  Entrez.email = "kstjohnamnh.org" #Replace
  with your own!
  ```

bioPython

- A module for biologists (based on bioPerl).
  bioPython page

- Need to install:
  `conda install biopython`

- Many useful functions but will focus on sequence I/O and interactions with NCBI (Entrez).
  `from Bio import Entrez`
  `#NCBI requires an email from each user:`
  `Entrez.email = "kstjohnamnh.org" #Replace`
  `with your own!`

- List all the databases we can access:

bioPython

- A module for biologists (based on bioPerl).
  bioPython page
- Need to install:
  `conda install biopython`
- Many useful functions but will focus on sequence I/O and interactions with NCBI (Entrez).
  ```
  from Bio import Entrez
  #NCBI requires an email from each user:
  Entrez.email = "kstjohnamnh.org" #Replace
  with your own!
  ```
- List all the databases we can access:`handle =
  Entrez.einfo()`
  ```
  record = Entrez.read(handle)
  record["DbList"]
  ```
- Can access individual databases, and records within (more in lab!)

seqIO

- Useful package for sequence input and output:

seqIO

- Useful package for sequence input and output:
  `from Bio import SeqIO`

seqIO

- Useful package for sequence input and output:
  `from Bio import SeqIO`

- Example (from biopython.org tutorial:)

```
from Bio import SeqIO
shortSequences = [] # Setup an empty list
for record in SeqIO.parse(open("cor6_6.gb", "rU"), "genbank"):

    if len(record.seq) < 300:
        # Add this record to our list
        shortSequences.append(record)

print "Found {0} short sequences".format(len(shortSequences))

outputHandle = open("shortSeqs.fasta", "w")
SeqIO.write(shortSequences, outputHandle, "fasta")
outputHandle.close()
```

Recap

- Using `bioPython` in lab today.

Recap

- Using `bioPython` in lab today.
- Email lab reports to `kstjohn@amnh.org`

Recap

- Using `bioPython` in lab today.
- Email lab reports to `kstjohn@amnh.org`
- Challenges available at `rosalind.info`