# Algorithmic Approaches for Biological Data, Lecture #14

Katherine St. John

City University of New York American Museum of Natural History

9 March 2016





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



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



- 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



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



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



- 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.



- 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.

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





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



- 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 ">".



- 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.



- 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.



- 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



- 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:



- 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).

```
where class-bgdarh-and wilgon-top* style-vertical-align: top> an none-'extoclibili.0.1.0.1.2" [da'extoclibili.0.1.0.1.2.2.2" [da'extoclibili.0.1.0.1.2.2.2" [da'extoclibili.0.1.0.1.2.2.2" [da'extoclibili.0.1.0.1.2.2" [da'extoclibili.0.1.0.1.2.2" [da'extoclibili.0.1.0.1.2.2" [da'extoclibili.0.1.0.1.2.2" [da'extoclibili.0.1.0.1.2.2" [da'extoclibili.0.1.0.1.2.2" [da'extoclibili.0.1.0.1.2.2" [da'extoclibili.0.1.0.2.2" [da'extoclibili.0.1.0.2.2.2" [da'extoclibili.0.1.0.2.2" [da'extoclibili.0.1.0.2.2" [da'extoclibili.0.1.0.2.2" [da'extoclibili.0.1.0.2.2" [da'extoclibili.0.1.0.2" [da'extoclibili.0.
```

• Including the module:

```
where class-bgdarh-and wilgon-top* style-vertical-align: top> an none-'extoclibili.0.1.0.1.2" [da'extoclibili.0.1.0.1.2.2.2" [da'extoclibili.0.1.0.1.2.2.2" [da'extoclibili.0.1.0.1.2.2.2" [da'extoclibili.0.1.0.1.2.2" [da'extoclibili.0.1.0.1.2.2" [da'extoclibili.0.1.0.1.2.2" [da'extoclibili.0.1.0.1.2.2" [da'extoclibili.0.1.0.1.2.2" [da'extoclibili.0.1.0.1.2.2" [da'extoclibili.0.1.0.1.2.2" [da'extoclibili.0.1.0.2.2" [da'extoclibili.0.1.0.2.2.2" [da'extoclibili.0.1.0.2.2" [da'extoclibili.0.1.0.2.2" [da'extoclibili.0.1.0.2.2" [da'extoclibili.0.1.0.2.2" [da'extoclibili.0.1.0.2" [da'extoclibili.0.
```

• Including the module: import urllib2

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

6/9

```
v/n-re class-hgs/driv-cd vilgon*top* style=vertical-align: top> a none=*etcod10013.0.1.0.1.2" (do*etcod10013.0.1.0.1.2">
v/n-re class-hgs/driv-cd vilgon*top* style=vertical-align: top> 8 mbsp;

v/n-re class-hgs/driv-cd vilgon*top* style=vertical-align: top> a none=*etcod10013.0.1.0.1.3" (do*etcod10013.0.1.0.1.3">
v/n-re class-hgs/driv-cd vilgon*top* style=vertical-align: top> Natthew Aurdens, American Museum of Natural History

v/n-re class-hgs/driv-cd vilgon*top* style=vertical-align: top> Natural selection in Lepidoptera: biological studies coress scales/quot;

v/n-re class-hgs/driv-cd vilgon*top* style=vertical-align: top> & mone=*etcod10013.0.1.0.1.4" (do*etcod10013.0.1.0.1.4">
v/n-re class-hgs/driv-cd vilgon*top* style=vertical-align: top> a none=*etcod10013.0.1.0.1.4" (do*etcod10013.0.1.0.1.4">
v/n-re class-hgs/driv-cd vilgon*top* style=vertical-align: top> a none=*etcod10013.0.1.0.1.4" (do*etcod10013.0.1.0.1.4">
v/n-re class-hgs/driv-cd vilgon*top* style=vertical-align: top> a none=*etcod10013.0.1.0.1.4"

v/n-re class-hgs/driv-cd vilgon*top* style=vertical-align: top> **
v/n-re
```

- Including the module: import urllib2
- Opening a file:
  - webpage =

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

```
Annual class-"bgdark"-red valigm"top" style="vertical-align: top" on name="etrodiB013,0.1.0.1.2" (da"etrodiB013,0.1.0.1.2" v/a-abb/Spring Term 2006/nb/dtb

Aligned Class-"bglight"-red valigm="top" style="vertical-align: top" on name="etrodiB013,0.1.0.1.2" (da"etrodiB013,0.1.0.1.2" v/a-abb/Spring Term 2006/nb/dtb

Aligned Class-"bglight"-red valigm="top" style="vertical-align: top" of name="etrodiB013,0.1.0.1.3" (da"etrodiB013,0.1.0.1.3" v/a-abb/Sp-Jan-16v/hb/dtb

Aligned Class-"bglight"-red valigm="top" style="vertical-align: top" of name="etrodiB013,0.1.0.1.4" (da"etrodiB013,0.1.0.1.4" v/a-abb/Sp-Jan-16v/hb/dtb

Aligned Class-"bglight"-red valigm="top" style="vertical-align: top" of name="etrodiB013,0.1.0.1.4" (da"etrodiB013,0.1.0.1.4" v/a-abb/Sp-Jan-16v/hb/dtb

Aligned Class-"bglight"-red valigm="top" style="vertical-align: top" of name="etrodiB013,0.1.0.1.4" (da"etrodiB013,0.1.0.1.4" v/a-abb/Sp-Jan-16v/hb/dtb

Aligned Class-"bglight"-red valigm="top" style="vertical-align: top" of name="etrodiB013,0.1.0.1.4" (da"etrodiB013,0.1.0.1.4" v/a-abb/Sp-Jan-16v/hb/dtb

Aligned Class-"bglight"-red valigm="top" style="vertical-align: top" of name="etrodiB013,0.1.0.1.4" (da"etrodiB013,0.1.0.1.4" v/a-abb/Sp-Jan-16v/hb/dtb

Aligned Class-"bglight"-red valigm="top" style="vertical-align: top" of name="etrodiB013,0.1.0.1.4" (da"etrodiB013,0.1.0.1.4" v/a-abb/Sp-Jan-16v/hb/dtb

Aligned Class-"bglight"-red valigm="top" style="vertical-align: top" of name="etrodiB013,0.1.0.1.4" (da"etrodiB013,0.1.0.1.4" v/a-abb/Sp-Jan-16v/hb/dtb

Aligned Class-"bglight"-red valign="top" style="vertical-align: top" of name="etrodiB013,0.1.0.1.4" (da"etrodiB013,0.1.0.1.4" v/a-abb/Sp-Jan-16v/hb/dtb

Aligned Class-"bglight"-red valign="top" style="vertical-align: top" of name="etrodiB013,0.1.0.1.4" (da"etrodiB013,0.1.0.1.4" v/a-abb/Sp-Jan-16v/hb/dtb

Aligned Class-"bglight"-red valign="top" style="vertical-align: top" of name="etrodiB013,0.1.0.1.4" v/a-abb/Sp-Jan-16v/hb/dtb

Aligned Class-"bglight "red valign="top" style="vertical-align: top" of na
```

- 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:

```
Anthor class-"galant"-act valign="top" style="vertical-align: top" on name="extodibili3.8.1.8.1.2" (de"extodibili3.8.1.8.1.2">v/a-vibi/Spring Term 2060/No/Vibi/
Arthor class-"galant"-act valign="top" style="vertical-align: top" begg;

Vibi/
Arthor class-"galant"-act valign="top" style="vertical-align: top" an name="extodibili3.8.1.8.1.3">v/a-vibi/Sp-3an-160/No/Vibi/
Arthor class-"galant"-act valign="top" style="vertical-align: top" an name="extodibili3.8.1.8.1.3">v/a-vibi/Sp-3an-160/No/Vibi/
Arthor class-"galant"-act valign="top" style="vertical-align: top" "Returned selection in Lepidoptero: biological studies across scalesdapot;

Arthor class-"galant"-act valign="top" style="vertical-align: top" beggs;

Arthor class-"galant"-act valign="top" style="vertical-align: top" beggs;

Arthor class-"galant"-act valign="top" style="vertical-align: top" beggs;

Arthor class-"galant"-act valign="top" style="vertical-align: top" beggs.

Arthor class-"galant beggs.

Arthor class-
```

- 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.

```
content class-"galari-red valign="top" style="vertical-align: top" on onee-"errock10013.0.1.0.1.2" (da"errock10013.0.1.0.1.2" via-velock5pring Term 2006/nbc/via-virted class-"galari-red valign="top" style="vertical-align: top" begin varied class-"galari-red valign="top" style="vertical-align: top" begin varied class-"galari-red valign="top" style="vertical-align: top" between class-"galari-red valign="top" style="vertical-align: top" between class-"galari-red valign="top" style="vertical-align: top" between class-"galari-red valign="top" style="vertical-align: top" below class-"galari-red valign="top" style="
```

- Including the module: import urllib2
- Opening a file:
  - webpage =

 ${\tt urllib2.urlopen("http://www.amnh.org/our-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-gilder-gilder-gilder-gilder-gilder-gilder-gilder-gilder-gilder-gilder-gilder-gilder-gilder-gilder-gilder-gilder-gilder-gilder-gilde$ 

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

```
denote class-logismi-set valign="top" style="vertical-align: top" on onee-"erocd10013,0.1.0.1.2" (de"erocd10013,0.1.0.1.2" /cd-add-Spring Term 2016/nfo-/rds

virsadr class-logismi-set valign="top" style="vertical-align: top" on onee-"erocd10013,0.1.0.1.2" (de"erocd10013,0.1.0.1.2" /cd-add-25-lon-16/nfo-/rds

virsadr class-logismi-set valign="top" style="vertical-align: top" on onee-"erocd10013,0.1.0.1.3" (de"erocd10013,0.1.0.1.3" /cd-arocd10013,0.1.0.1.3" /cd-arocd10013,0.1.0.1.3" /cd-arocd10013,0.1.0.1.3" /cd-arocd10013,0.1.0.1.3" /cd-arocd10013,0.1.0.1.4" /cd-arocd10013,0.1.0.1.4" (de"erocd10013,0.1.0.1.4" /cd-arocd10013,0.1.0.1.4" /cd-arocd10013,0.1.0.1.4"
```

- Including the module: import urllib2
- Opening a file:
  - webpage =

 ${\tt urllib2.urlopen("http://www.amnh.org/our-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-gilder-graduate-school/academics-gilder-graduate-gilder-gild$ 

- 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.

```
consent class-"biglight">

consent class-"biglight consent class-"biglight consent class-"biglight consent class-"biglight consen
```

- Including the module: import urllib2
- Opening a file:
- webpage =

 ${\tt urllib2.urlopen("http://www.amnh.org/our-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-gilder-gilder-gilder-gilder-gilder-gilder-gilder-gilder-gilder-gilder-gilder-gilder-gilder-gilder-gilder-gilder-gilder-gilder-gilde$ 

- 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:

```
where class-"gates"-act valige "top" style-"vertical-align: top" on none-"extendingl.8.1.8.1.2" (d-"extendingl.8.1.8.1.2"-vio-de-Spring Term 2016-/fe-vio-de-Spring Term 2
```

- Including the module: import urllib2
- Opening a file:
- webpage =

 ${\tt urllib2.urlopen("http://www.amnh.org/our-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-school/academics-and-research/richard-gilder-graduate-gilder-gilder-gilder-gilder-gilder-gilder-gilder-gilder-gilder-gilder-gilder-gilder-gilder-gilder-gilder-gilder-gilder-gilder-gilde$ 

- 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()

• A module for biologists (based on bioPerl).



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





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

7 / 9



- 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).

7 / 9



- 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!



- 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:



- 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()</pre>
```

# 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