BST281: Genomic Data Manipulation, Spring 2018

Monday 07: Python data input and output

Python organizes local variables into blocks, and all variables into namespaces.

To use variables (or functions) from another module, use the **import** command.

Many useful modules are built into Python: <http://docs.python.org>

The most common module is **sys**: it links your program up to the outside world, i.e. system.

**import sys** to access.

**sys.argv** is a list of strings that contain any command line parameters given to script at runtime.

**sys.argv[0]** is always the file name of your Python script.

**python script.py one 2 three** makes **sys.argv = ["script.py", "one", "2", "three"]**.

IO (input and output) is performed in Python using streams.

Cursors that read data from a location (and advance) or write data to a location (and advance).

**sys.stdin** and **sys.stdout** are built-in standard input and output streams.

Come from the keyboard and to the screen by default; can be redirected using **<** and **>**.

**python script.py < input.txt > output.txt**

**sys.stderr** is an unbuffered output stream useful for debugging; separately redirectable using **2>**.

File IO streams can be created as variables using **open**:

**istm = open( "input.txt", "r" )**

**strInput = istm.readline( )**

**ostm = open( "output.txt", "w" )**

**ostm.write( strOutput )**

Don't forget to **close** IO streams (especially output).

Useful constructs:

**with open( strFilename ) as istm:**

**for strLine in istm:**

For delimited (typically tab, sometimes comma) files, you can import the built-in **csv** module:

**for astrLine in csv.reader( istm, csv.excel\_tab ):**

**strColumn1, strColumn2, strColumn3 = astrLine**

**with csv.writer( ostm, csv.excel\_tab ) as csvw:**

**csvw.writerow( [strString, iInt, dFloat] )**

Beware of mismatched end-of-line characters (\r and \n) and use text mode carefully.

Beware of accidentally overwriting files using **open( strOutput, "w" )** or **>**.

# Textbooks

Python I/O: Haddock and Dunn, Chapter 10 p173-184, 188-192, Chapter 11 p201-214

# Literature

[Confronting the ethical challenges of big data in public health. Bourne, PLoS Comp Bio 2015](https://www.ncbi.nlm.nih.gov/pubmed/25664660)

[The Virtuous Cycle of a Data Ecosystem. Voytek, PLoS Comp Bio 2016](https://www.ncbi.nlm.nih.gov/pubmed/27490108)